

DOCUMENT RESUME

ED 196 099

CS 503 206

TITLE Status Report on Speech Research: A Report on the Status and Progress of Studies on the Nature of Speech, Instrumentation for Its Investigation, and Practical Applications, April 1-June 30, 1980.

INSTITUTION Haskins Labs., New Haven, Conn.

REPORT NO SR-62 (1980)

PUB DATE 80

NOTE 240p.

EDRS PRICE MF01/PC10 Plus Postage.

DESCRIPTORS *Acoustic Phonetics; *Applied Linguistics; *Articulation (Speech); *Auditory Perception; *Communication Research; Consonants; *Oral Language; Pronunciation; Speech Handicaps; Speech Skills; Vowels

ABSTRACT

Research reports on the nature of speech, instrumentation for the investigation of speech, and practical applications of research are included in this status report for the April 1-June 30, 1980, period. The reports deal with the following topics: (1) the perceptual equivalent of two acoustic cues for a speech contrast is specific to phonetic perception; (2) duplex perception of acoustic patterns as speech and nonspeech; (3) evidence for phonetic processing of cues to place of articulation; (4) some articulatory correlates of perceptual isochrony; (5) effects of utterance continuity on phonetic judgments; (6) laryngeal adjustments in stuttering; (7) letter detection errors on word endings; (8) speaking rate, syllable stress, and vowel identity; (9) sonority and syllabicity as acoustic correlates of perception; (10) influence of vocalic context on perception of the "sh"/"s" distinction; (11) the role of auditory memory in speech perception and discrimination; (12) talker-independent information for vowel identity; (13) context-conditioned specification of vowel identity; and (14) influence of preceding fricative on stop consonant perception. (FL)

 * Reproductions supplied by EDRS are the best that can be made *
 * from the original document. *

ED196099

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

SR-62 (1980)

Status Report on

SPEECH RESEARCH

A Report on
the Status and Progress of Studies on
the Nature of Speech, Instrumentation
for its Investigation, and Practical
Applications

1 April - 30 June 1980

Haskins Laboratories
270 Crown Street
New Haven, Conn. 06510

Distribution of this document is unlimited.

(This document contains no information not freely available to the general public. Haskins Laboratories distributes it primarily for library use. Copies are available from the National Technical Information Service or the ERIC Document Reproduction Service. See the Appendix for order numbers of previous Status Reports.)

CS 503 206

ACKNOWLEDGMENTS

The research reported here was made possible in part by support from the following sources:

National Institute of Child Health and Human Development
Grant HD-01994

National Institute of Child Health and Human Development
Contract N01-HD-1-2420

National Institutes of Health
Biomedical Research Support Grant RR-05596

National Science Foundation
Grant MCS79-16177
Grant BNS78-27331

National Institute of Neurological and Communicative
Disorders and Stroke
Grant NS13870
Grant NS13617

National Institute of Arthritis, Metabolism, and
Digestive Diseases
Grant AM25814

HASKINS LABORATORIESPersonnel in Speech Research

Alvin M. Liberman,* President and Research Director
 Franklin S. Cooper,* Associate Research Director
 Patrick W. Nye, Associate Research Director
 Raymond C. Huey, Treasurer
 Alice Dadourian, Secretary

Investigators

Arthur S. Abramson*
 Thomas Baer
 Fredericka Bell-Berti+
 Catherine Best+
 Gloria J. Borden*
 Susan Brady*
 Guy Carden*
 Robert Crowder*
 William Ewan*
 Carol A. Fowler*
 Jane H. Gaitenby
 Katherine S. Harris*
 Alice Healy*
 Kiyoshi Honda¹
 Leonard Katz*
 Scott Kelso
 Andrea G. Levitt*
 Isabelle Y. Liberman*
 Leigh Lisker*
 Anders Løfqvist²
 Virginia Mann+
 Charles Marshall
 Ignatius G. Mattingly*
 Nancy S. McGarr*
 Lawrence J. Raphael*
 Bruno H. Repp
 Philip E. Rubin
 Donald P. Shankweiler*
 Michael Studdert-Kennedy*
 Michael T. Turvey*
 Robert Verbrugge*
 Hirohide Yoshioka¹

Technical and Support Staff

Eric L. Andreasson
 Elizabeth P. Clark
 Vincent Gulisano
 Donald Hailey
 Terry Halwes
 Sabina D. Koroluk
 Agnes M. McKeon
 Nancy O'Brien
 Marilyn Parnell
 William P. Scully
 Richard S. Sharkany
 Leonard Szubowicz
 Edward R. Wiley
 David Zeichner

Students*

David Dechovitz
 Laurie Feldman
 Hollis Fitch
 Carole E. Gelfer
 David Goodman
 Janette Henderson
 Charles Hoequist
 Kenneth Holt
 Robert Katz
 Aleksandar Kostic
 Peter Kugler
 Anthony Levas
 Harriet Magen
 Suzi Pollack
 Patti Jo Price
 Sandra Prindle
 Brad Rakerd
 Arnold Shapiro
 Louis G. Tassinari
 Janet Titchener
 Emily Tobey-Cullen
 Betty Tuller
 N. S. Viswanath
 Douglas Whalen
 Deborah Wilkenfeld

*Part-time

¹Visiting from University of Tokyo, Japan

²Visiting from Lund University, Sweden

+NIH Research Fellow

CONTENTSI. Manuscripts and Extended Reports

The Perceptual Equivalence of Two Acoustic Cues for a Speech Contrast Is Specific to Phonetic Perception--Catherine T. Best, Barbara Morrongiello and Rick Robson	 1
Duplex Perception of Acoustic Patterns as Speech and Nonspeech--Alvin M. Liberman and David S. Isenberg	 47
Evidence for Phonetic Processing of Cues to Place of Articulation: Perceived Manner Affects Perceived Place--Guy Carden, Andrea Levitt, Peter W. Jusczyk and Amanda Walley	 59
Some Articulatory Correlates of Perceptual Isochrony--Betty Tuller and Carol A. Fowler	 87
Effects of Utterance Continuity on Phonetic Judgments--David R. Dechovitz, Brad Rakerd and Robert R. Verbrugge	 101
Laryngeal Adjustments in Stuttering: A Glottographic Observation Using a Modified Reaction Paradigm--Hirohide Yoshioka and Anders Lofqvist	 117
Missing -ing in Reading: Letter Detection Errors on Word Endings--Adam Drewnowski and Alice F. Healy	 127
Speaking Rate, Syllable Stress, and Vowel Identity--Brad Rakerd, Robert R. Verbrugge and Donald P. Shankweiler	 149
Sonority and Syllabicity: Acoustic Correlates of Perception--Patti Jo Price	 161
Influence of Vocalic Context on Perception of the [ʃ]-[s] Distinction: III. Acoustic Observations--Virginia A. Mann and Bruno H. Repp	 177
The Role of Auditory Memory in Speech Perception and Discrimination--Robert G. Crowder	 187
Talker-Independent Information for Vowel Identity--Robert R. Verbrugge and Brad Rakerd	 205
Context-Conditioned Specification of Vowel Identity--Robert R. Verbrugge, Donald P. Shankweiler and Carol A. Fowler	 217
Influence of Preceding Fricative on Stop Consonant Perception: Some Amendments--Bruno H. Repp and Virginia A. Mann	 225

II. Publications

..... 239

III. Appendix: DTIC and ERIC numbers (SR-21/22 - SR-61)

..... 241

SR-62 (1980)

I. MANUSCRIPTS AND EXTENDED REPORTS

THE PERCEPTUAL EQUIVALENCE OF TWO ACOUSTIC CUES FOR A SPEECH CONTRAST IS SPECIFIC TO PHONETIC PERCEPTION*

Catherine T. Best, Barbara Morrongiello+ and Rick Robson+

Abstract. Trading relations show that diverse acoustic consequences of minimal contrasts in speech are equivalent in perception of phonetic categories. This perceptual equivalence received stronger support from a recent finding that discrimination was differentially affected by the phonetic cooperation or conflict between two cues for the /slit/-/split/ contrast. Experiment 1 extended the trading relation/perceptual equivalence findings to the /sei/-/stei/ contrast. With a more sensitive discrimination test, Experiment 2 found that cue equivalence is a characteristic of perceptual sensitivity to phonetic information. Using "sinewave analogues" of the /sei/-/stei/ stimuli, Experiment 3 showed that perceptual integration of the cues was phonetic, not psychoacoustic, in origin. Only subjects who perceived the sinewave stimuli as "say" and "stay" showed a trading relation and perceptual equivalence; subjects who perceived them as nonspeech failed to integrate the two dimensions perceptually. Moreover, the pattern of differences between obtained and predicted discrimination was quite similar across the first two experiments and the "say"- "stay" group of Experiment 3. The pattern suggested that phonetic perception was responsible for better-than-predicted performance by these groups. Trading relations and perceptual equivalence thus appear to derive specifically from perception of phonetic information.

INTRODUCTION

Research with a variety of minimal segmental distinctions in synthetic speech has shown that perception of a given contrast can be cued by

*Revised version of a paper presented at the 98th Meeting of the Acoustical Society of America, Salt Lake City, Utah (November, 1979). Throughout this paper, the formants are abbreviated as F1, F2, F3, and the fundamental as F0. A slightly revised version of this paper is to appear in Perception & Psychophysics, in press.

+University of Massachusetts, Amherst

Acknowledgment: This work was supported by NICHD Grant #HD01994 to the Haskins Laboratories and NINCDS Postdoctoral Fellowship Grant #NS5085 to the first author. We thank the following people for their valuable contributions: Alvin M. Liberman, for use of the Haskins facilities, and for helpful advice throughout the project; Terry G. Halwes, for instruction on stimulus development; and Michael Studdert-Kennedy, Bruno H. Repp, Robert E. Remez, Robert Verbrugge, and Steven S. Braddon, for helpful discussions and suggestions on earlier drafts of this paper.

appropriate change in the major acoustic property that differentiates the phonetic contrast in natural speech (e.g., Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Liberman, Harris, Hoffman, & Griffith, 1961; Liberman & Studdert-Kennedy, 1976; Studdert-Kennedy, 1976). However, minimal articulatory contrasts result in concurrent differences along more than one acoustic dimension. As this last fact might suggest, perceptual studies of multiple cues for phonetic contrasts indicate that listeners make use of the various acoustic consequences of spoken segmental distinctions. For example, voicing distinctions can be cued not only by changes in voice onset time (VOT—as an acoustic measure), with all else held constant, but also by changes in F1 onset frequency, F0 contour, or aspiration energy (e.g., Haggard, Ambler, & Callow, 1970; Lisker, 1975, 1978; Lisker, Liberman, Erickson, Dechovitz, & Mandler, 1977; Repp, 1979), all of which are acoustic correlates of laryngeal timing distinctions in stop production (Abramson & Lisker, 1965). A variety of acoustic consequences of articulatory distinctions have also been found to serve as cues for place (e.g., Dorman, Studdert-Kennedy, & Raphael, 1977; Harris, Hoffman, Liberman, Delattre, & Cooper, 1958) and manner of articulation (e.g., Dorman, Raphael, & Liberman, 1979; Miller & Liberman, 1979; Repp, Liberman, Eccardt, & Pesetsky, 1978).

The fact that various acoustic properties serve as cues for a speech contrast suggests that, in a functional sense, those diverse cues provide equivalent information about the phonetic category distinction involved (cf. Dorman et al., 1977; Repp et al., 1978). If different cues do provide equivalent phonetic information, we would expect to find perceptual trading relations among them. That is, it should be possible to offset a "weakness" in one cue by strengthening the value of another (within limits, which are possibly defined by the acoustic effects of natural articulations). Empirical work has supported this hypothesis—trading relations have been found among diverse cues for voicing (e.g., Summerfield & Haggard, 1977), place (e.g., Bailey & Summerfield, 1978; Hoffman, 1958), and manner distinctions (e.g., Dorman, Raphael, & Isenberg, 1979). However, while trading relations indicate that different acoustic cues may provide functionally equivalent information about phonetic categories, they cannot support a stronger claim that the cues yield qualitatively equivalent percepts. Trading relation studies have typically used forced-choice identification tests, which merely assess whether any of several acoustic manipulations are acceptable as cues for a given contrast. Forced-choice tests do not measure whether the same-category percepts based on different acoustic cues are identical in quality. Complete equivalence in the perceptual qualities of phonetic cues will be referred to as perceptual equivalence.

In a recent experiment on the /slIt/-/splIt/ contrast, Fitch, Halwes, Erickson, and Liberman (1980) conducted a more stringent test of perceptual equivalence between two cues for the medial stop /p/. A trading relation between two synthetic /slIt/-/splIt/ continua showed that when the formant transitions following silent closure were appropriate for a natural "slit" (/slIt/-biased continuum), listeners needed a significantly longer closure gap than when the transitions were appropriate for "split" (/splIt/-biased continuum), in order to hear "split" 50% or more of the time. Thus, additional silence compensated perceptually for "weakness" of the /slIt/-biased spectral cue. If the two cues had converged upon a unitary speech percept ("split") because of their common origin in stop consonant articulation, Fitch et

al. (1980) reasoned, differently-cued stimuli should be difficult to discriminate within a phonetic category (i.e., they should be perceptually equivalent).

Mere demonstration of poor within-category discriminability between cues would not provide convincing support for perceptual equivalence, however, since the null hypothesis cannot be proven. Therefore, Fitch et al. tested whether discrimination performance would be differentially affected by cooperation or conflict of the two cues along the phonetic dimension, using an oddity procedure that included three types of comparisons between stimuli from the two continua. In "two conflicting cues" comparisons, /slIt/-biased stimuli (spectral bias toward "slit") had longer closures (temporal bias toward "split") than /splIt/-biased stimuli, such that the two cues exactly cancelled one another phonetically. In the "two cooperating cues" comparisons, the two cues complemented each other phonetically--on all trials the /splIt/-biased stimuli had a longer closure gap (by the same amount of difference as in "two conflicting cues") than the /slIt/-biased stimuli. In "one cue" comparisons, the stimuli contrasted only on the spectral dimension.

The "phonetic" hypothesis was that if the two cues showed perceptual equivalence along a single phonetic dimension, /slIt/-biased and /splIt/-biased stimuli would be quite difficult to discriminate when they belonged to the same phonetic category, as would be the case for all "conflicting cues" comparisons. In contrast, they should be comparatively easy to discriminate when they belonged to different phonetic categories; this would be the case for those "one cue" comparisons that straddled the category boundary. Enhancing the between-category differences should lead to the highest discrimination performance; this was accomplished by those "cooperating cues" comparisons that straddled the phonetic category boundary. The alternative "auditory" hypothesis was that the two cues might remain discriminable on an auditory basis. In that case, performance would be equally high across the board for both "two cue" comparison types, since they contrasted along two acoustic dimensions, relative to performance on "one cue" comparisons, which contrasted on only one acoustic dimension. The results clearly supported the "phonetic" hypothesis, indicating that the two acoustic cues were perceptually equivalent along a single dimension in speech.

The 3-way oddity results may offer an important contribution to our knowledge about the conditions under which information from diverse acoustic dimensions is integrated in speech perception. However, the extent and reliability of perceptual equivalence among phonetic cues still needs assessment. To test the generality of the findings, other acoustic cues and phonetic category contrasts should be explored; indeed, the many reported trading relations make it unlikely that perceptual cue equivalence would be idiosyncratic to /slIt/-/splIt/. Experiment 1 of this paper extended the trading relation/perceptual equivalence concepts to the /sei/-/stei/ contrast,¹ through forced-choice identification and 3-way oddity discrimination tests with two /sei/-/stei/ continua that incorporated both spectral (F1 onset frequency) and temporal (silent closure duration) variation. The /sei/-/stei/ contrast is simpler than /slIt/-/splIt/ in phonetic, articulatory, and acoustic properties. Because of the nearly identical place of articulation for /s/ and /t/, /sei/ and /stei/ are dynamically similar. Each starts with the tongue edges pressed against the inner sides of the upper teeth, tongue-

tip in contact (or nearly so) with the alveolar ridge and/or the inner side of the juxtaposed front teeth (for the fricative /s/), and each ends with a more open vocal tract (for the diphthong /ei/). The result is very similar formant transitions in the vocalic portions of the two words. The major acoustic consequences of complete linguo-alveolar (or -dental) closure for a /t/ between /s/ and /ei/ are the introduction of a silent gap and a lower F1 onset frequency. (For general accounts of stop closure properties, see Delattre, Liberman, & Cooper, 1955; Fant, 1962; Stevens, 1971, 1974.) In the /slIt/-/splIt/ contrast, on the other hand, bilabial juxtaposition and the consequent labial transitions in the upper formants occur only for /splIt/.

If wider support is found (in Experiment 1) for the suggestion that perceptual integration of diverse acoustic cues takes account of their common origin in speech production, then it may be that trading relations and perceptual equivalence between cues are specific to the perception of phonetic information. Experiments 2 and 3 were therefore designed to test two alternatives to the notion that such findings may be unique to phonetic perception.

First, the oddity procedure may not provide the optimal test for true equivalence of the perceptual qualities of phonetic cues. That procedure is widely believed to place heavy demands on auditory short term memory, which may have encouraged listeners to categorize each stimulus while it was being presented. As a result, listeners might have been distinguishing among categorizations of the stimuli, rather than discriminating more fine-grained perceptual qualities that might be available prior to categorization. Moreover, since the discrimination test was administered after the forced-choice identification test, test-order could also have introduced a bias toward categorizing stimuli before discriminating them. Experiment 2 attempted to surmount these potential problems by using a more sensitive discrimination test with lower memory demands, and by collecting the discrimination data prior to the identification data.

Second, although it has been suggested that perceptual equivalence between diverse speech cues would occur only for perception of phonetic category information (cf. Fitch et al., 1980), no direct studies have been conducted for trading relations and perceptual equivalence between acoustic cues in nonspeech sounds. Experiment 3 tested the "psychoacoustic" alternative that trading relation/perceptual equivalence findings might occur for nonspeech sounds with complex acoustic properties like those used in our /sei/-/stei/ contrasts.

EXPERIMENT 1

Method

Subjects

The subjects were 15 undergraduate students. Ten subjects from Yale University were tested at Haskins Laboratories, and received \$3/hr for participation. The other five were from the University of Massachusetts at Amherst, and received grade-credits toward their introductory psychology

courses; they completed the tests in an auditory perception laboratory at their psychology department. All subjects reported having normal hearing in both ears (no diagnosed hearing losses).

Stimuli

Two 290 msec, three-formant vocalic syllables were created on the Haskins parallel-resonance synthesizer. They were stylized versions of the vocalic portions from natural, male utterances of "say" and "stay," and differed from one another only in F1 onset frequency (230 Hz vs. 430 Hz), as the acoustic analyses in the left-hand panel of Figure 1 illustrate. The formant amplitudes and overall amplitude envelopes of the stimuli were identical, as were the frequency characteristics of F2 and F3, and also of F1 beyond the initial 40 msec transition differences (details in Appendix A). For comparison of these synthetic syllables to natural speech, acoustic analyses were made on five tokens each of "say" and "stay" uttered by two males (center and right-hand panels of Figure 1); these analyses showed that, as expected, the most pronounced spectral difference between the vocalic parts of "say" and "stay" was a lower F1 onset frequency for "stay." Acoustic analyses of natural utterances of "day" were also included, since "day" and the vocalic portion of "stay" involve nearly identical articulatory gestures.² As would be expected, based on their articulatory similarity, the vocalic portion of "stay" was virtually identical to "day" in formant frequency characteristics. Figure 2 shows spectrograms of a natural "say" and "stay" and of "stay" tokens made from the two synthetic syllables by preceding each with a natural /s/ and a silent (closure) interval.

To test whether listeners would identify the synthetic syllables as "day" (i.e., to determine how well their F1 onset frequencies would support perception of an alveolar stop, in the absence of an /s/ and a closure gap), 12 additional synthetic syllables were generated for an "ay"-"day" continuum, in which F1 onset was varied between 160 Hz and 611 Hz (the F1 steady-state) in 33 Hz steps. A randomized forced-choice identification test (10 judgments/token) with 18 naive listeners (9 from Yale, 9 from University of Massachusetts/Amherst) revealed a fairly sharp category boundary. The test syllable with the 430 Hz F1 onset was perceived as nearly equivocal between "ay" and "day," and will hereafter be called weak "day," abbreviated d. The stimulus with the 230 Hz F1 onset was identified 100% of the time as "day," and will be called strong "day," abbreviated D.

The D and d syllables were each used to generate a "say"-"stay" continuum that incorporated a natural 120-msec /s/ from a male "say" utterance. The /s/ and the synthetic syllable were separated by silent gaps ranging between 0-136 msec, in 8 msec increments, resulting in two "say"-"stay" continua with 18 members each (from s[0]d to s[136]d, and from s[0]D to s[136]D).

Procedure

For the forced-choice identification test, a randomized sequence of 360 single-item trials was generated, with 2.5 msec interstimulus intervals (ISIs). The sequence contained ten repetitions of all items from the two "say"-"stay" continua, and was presented in sound-attenuated test rooms at a

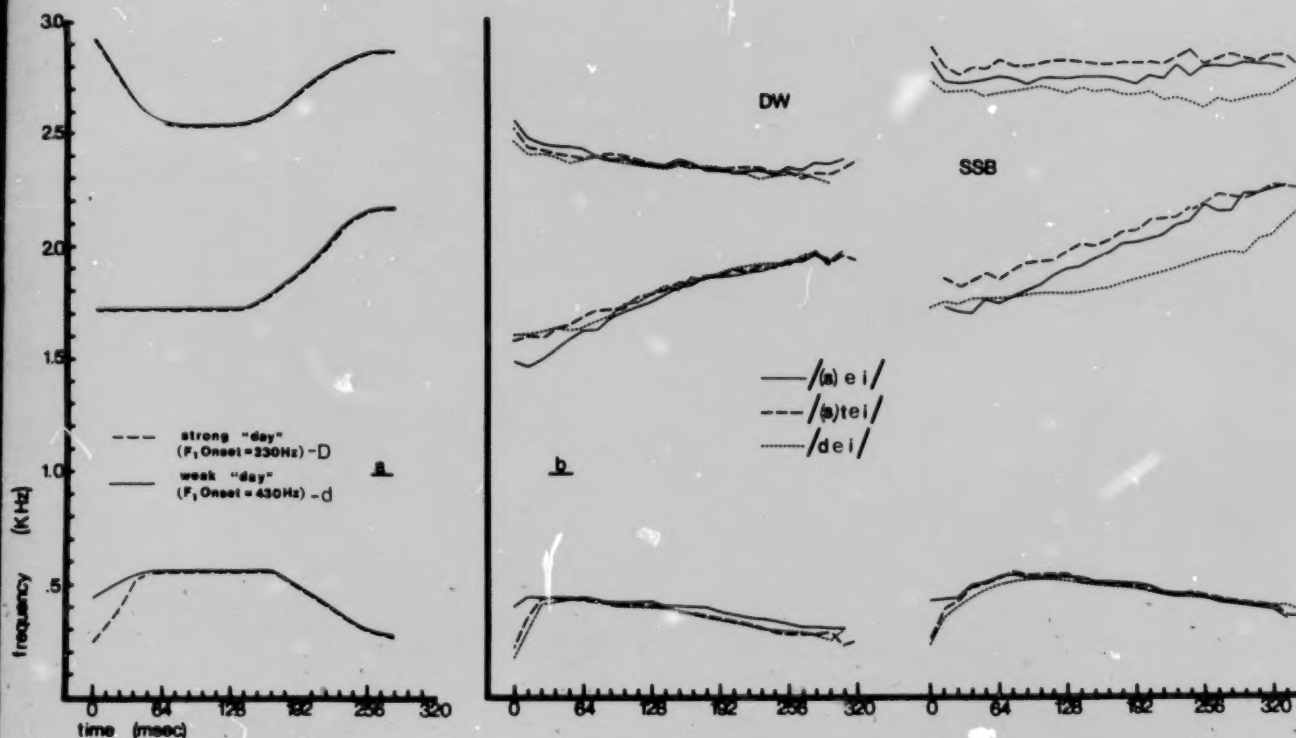


Figure 1. Acoustic measurements of F1, F2, and F3 for the following stimuli, from spectral sections of consecutive 12.8 msec windows: (a) synthetic weak "day" and strong "day," (b) averages of five tokens each of "say," "stay," and "day" by male talker DW (Fort Worth, Texas), and (c) by male talker SSB (Brooklyn, New York—variations in SSB's F2 and F3 frequencies reflect slight vowel color variation among these words in his dialect).

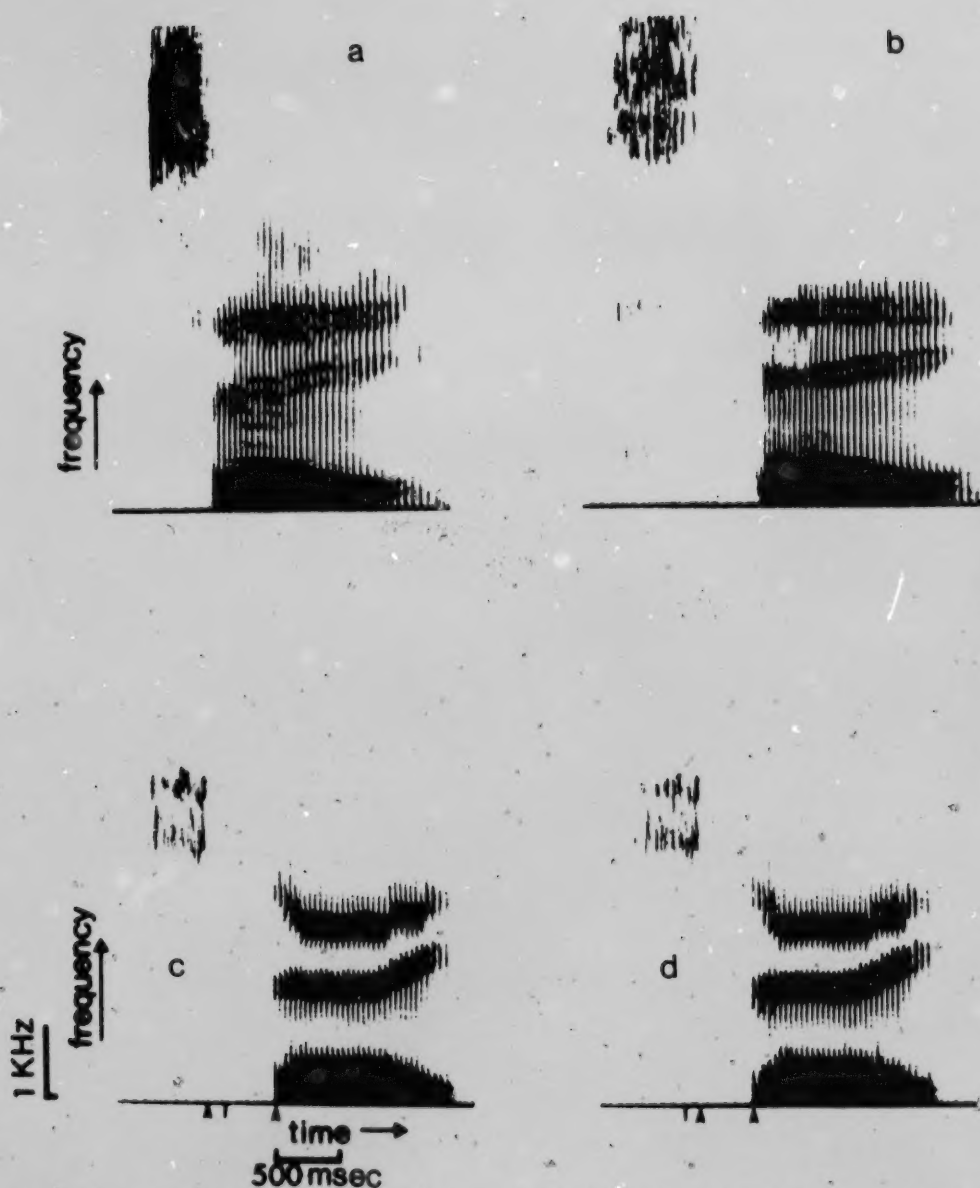


Figure 2. Spectrograms of natural, male utterances of (a) "say," and (b) "stay," by talker SSB, and synthesized vocalic tokens of "stay" made from (c) weak "day" and (d) strong "day." The synthetic tokens are preceded by a natural 120-msec /s/, and although they had different silent gap durations, a trading relation showed them to be identified equally consistently as "stay."

7

comfortable listening level (approximately 75 dB). Subjects identified each stimulus in writing as "say" or "stay."

Subjects took a 15-minute break following the identification test, and then completed a 3-way oddity discrimination test that included the following comparison types: "one cue," "two cooperating cues," and "two conflicting cues." In the 18 possible "one cue" comparisons, the three stimuli on each trial had identical gap durations (comparisons covered the entire 0-136 msec range), and the 'odd' stimulus differed from the other two in its F1 onset frequency (D vs. d). For both two-cue comparison types, a 24 msec gap difference was chosen to compensate phonetically for the F1 difference.³ In the 15 possible "cooperating cues" comparisons, D stimuli always had a 24 msec longer silent gap (both cues biased toward "stay") than d stimuli, so that the phonetic complementarity of the cues enhanced the between-category differences; comparisons ranged from s[0]d-s[24]D to s[112]d-s[136]D. For the 15 possible "conflicting cues" comparisons, d stimuli (spectral bias toward "say") had a 24 msec longer gap (temporal bias toward "stay") than D stimuli, so that phonetic cancellation between the cues minimized any between-category differences; these comparisons ranged from s[0]D-s[24]d to s[112]D-s[136]d. Discrimination should be facilitated for "cooperating cues" comparisons that straddled the category boundary (because between-category differences were enhanced for those comparisons), relative to "one cue" comparisons that straddled the boundary. Discrimination of the "conflicting cues" comparisons should be lowest, because between-category differences were minimized throughout the range of comparisons tested--the "conflicting cues" comparisons never straddled the boundary.

The discrimination test was a randomized sequence of all stimulus comparisons for all three comparison types, and included six presentations⁴ of each of the 48 possible stimulus comparisons (total items = 288). Within each trial, ISIs were 1 sec, and intertrial intervals (ITIs) were 3 sec.

Results and Discussion

Identification Test

The results for the forced-choice identification test are shown in Figure 3, with gap durations along the abscissa and percent "stay" responses along the ordinate. The mean category boundary (50% "stay" responses) for the strong "day" (D) function fell at 32.4 msec (range = 11.4 - 52.0 msec), and that for the weak "day" (d) function at 57.1 msec (range = 40.0 - 94.0 msec). The average magnitude of the trading relation between the two continua was thus 24.6 msec (the amount of silence that compensated for the phonetic effects of d vs. D). All subjects showed a category boundary at smaller gap durations for the D continuum than for the d continuum, although the magnitude of the difference varied (range = 9.3 - 54.0 msec). The mean difference between category boundaries was significant, $t = 7.23$, $p < .001$, supporting the expected trading relation. To be perceived as "stay," the d stimuli required approximately 24 msec more silence between the /s/ and the vocalic syllable than did the D stimuli.

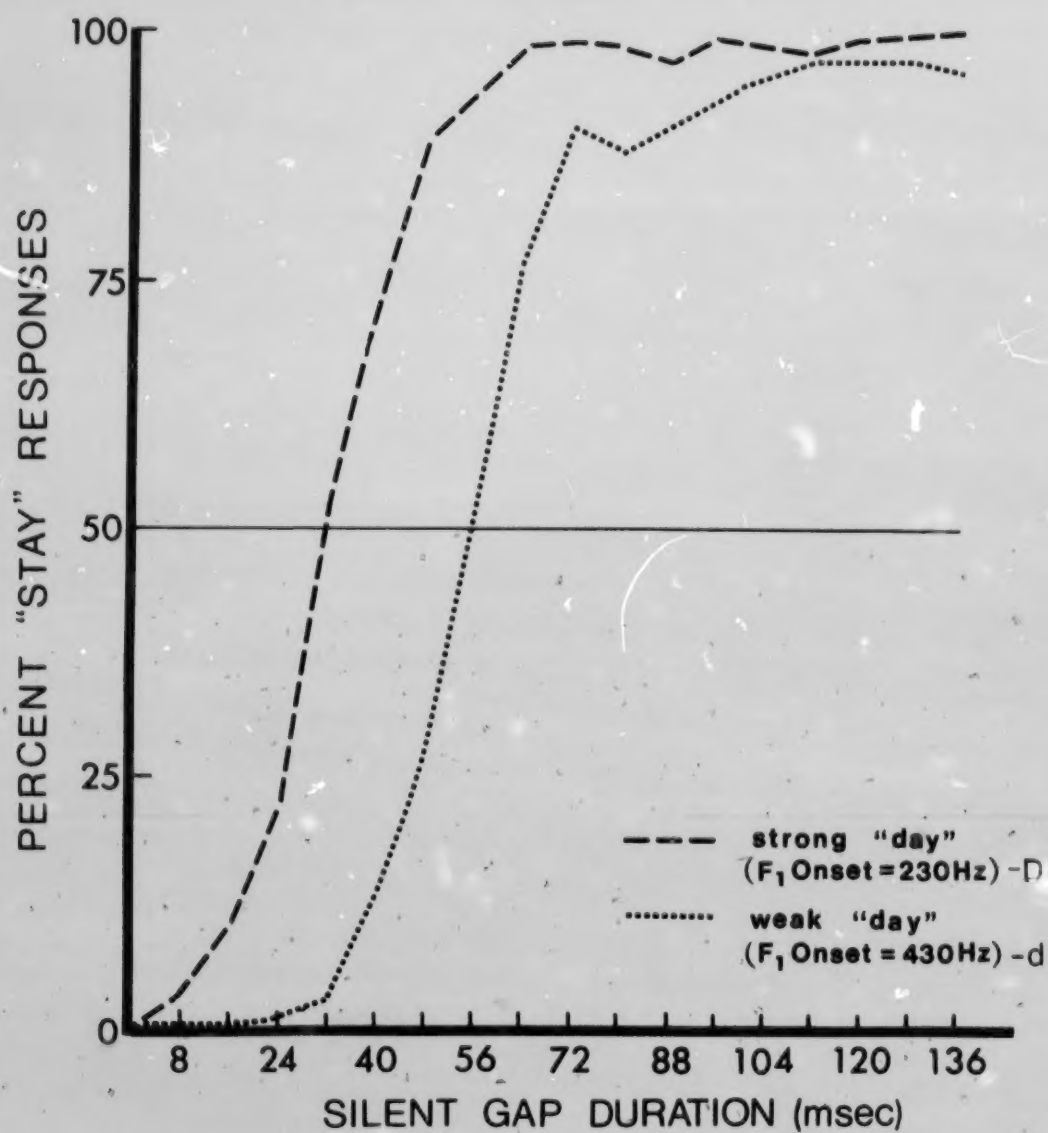


Figure 3. Identification functions for the strong "day" and weak "day" continua in Experiment 1.

Three-way Oddity Test

Group findings. The results for the 3-way oddity test are shown in Figure 4, with percent correct responses on the ordinate. Obtained functions for the three comparison types are represented in the left panel; the right panel represents the corresponding Predicted functions based on the identification data (formula in Appendix B). The Obtained data were submitted to an analysis of variance (ANOVA) crossing three Comparison Types with 15 Stimulus Pairs, which included only the stimulus comparisons in the range of overlap between "one cue" and two-cues comparison types (mean gap durations per comparison of 12-128 msec). The significant main effect of Comparison Types, $F(2,28) = 34.14$, $p < .001$, supported the perceptual equivalence predictions; the order of performance levels was "cooperating cues" > "one cue" > "conflicting cues." (Table 1 shows results of Tukey pairwise contrasts for this and subsequent Comparison Types effects.) Performance near the category boundary was significantly higher than within-category performance, according to the main effect of Stimulus Pairs, $F(14,196) = 7.01$, $p < .001$. Moreover, the Comparison Types x Stimulus Pairs interaction, $F(28,392) = 3.31$, $p < .001$ (which was broken down by simple effects tests), indicated that the magnitude of peak vs. trough level differences followed the order: "cooperating cues" ($F(14,588) = 27.13$, $p < .001$) > "one cue" ($F(14,588) = 2.39$, $p < .005$) > "conflicting cues" ($F(14,588) = 1.75$, $p = .05$).⁵

To determine whether the Obtained pattern derived from the trading relation in identification judgments, an ANOVA was also performed on the Predicted data for three Comparison Types x 15 Stimulus Pairs. The pattern of results from this ANOVA was essentially the same as for the Obtained data, as Figure 4 illustrates (Stimulus Pairs effect, $F(14,196) = 10.72$, $p < .001$; Comparison Types effect, $F(2,28) = 29.60$, $p < .001$ — see results of Tukey pairwise contrasts among Comparison Types, Table 1). Simple effects tests on the Comparison Types x Stimulus Pairs interaction, $F(28,392) = 33.13$, $p < .001$, were also highly similar to the Obtained results ("cooperating cues," $F(14,588) = 17.76$, $p < .001$; "one cue," $F(14,588) = 4.36$, $p < .001$; "conflicting cues," $F(14,588) = .22$, n.s.). Thus, the pattern of performance on the 3-way oddity test was predicted from the trading relation results on the identification test, suggesting that silent closure and F1 onset frequency are qualitatively equivalent along a single phonetic dimension.

Obtained vs. Predicted differences. Although the Obtained and Predicted data showed similar patterns, Figure 4 suggests that Obtained discrimination was better than Predicted, particularly beyond the immediate neighborhood of the performance "peaks" that had been expected for clear between-category comparisons. This observation was supported by ANOVAs for each of the three comparison types, which crossed the two Functions (Obtained vs. Predicted) with 15 Stimulus Pairs (12-128 msec range). The three significant Functions effects showed that Obtained performance was better than Predicted for "cooperating cues," $F(1,14) = 8.82$, $p < .025$, "one cue," $F(1,14) = 12.22$, $p < .005$, and "conflicting cues" comparisons, $F(1,14) = 19.39$, $p < .001$. In addition, significant Functions x Stimulus Pairs interactions for "one cue," $F(17,238) = 2.54$, $p < .001$, and "cooperating cues" comparisons, $F(17,238) = 4.56$, $p < .001$, revealed that Obtained performance was higher than

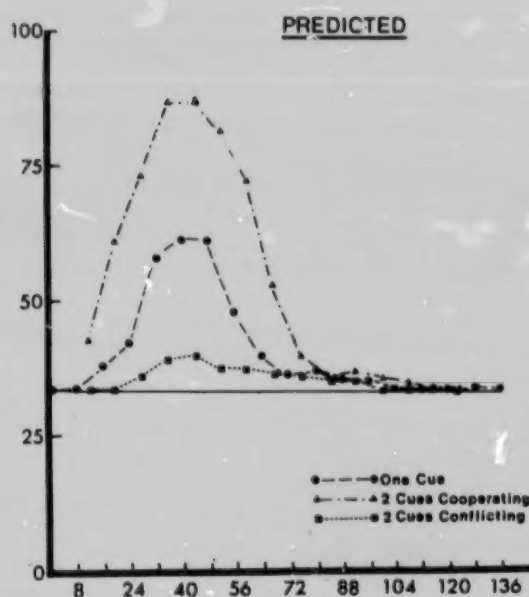
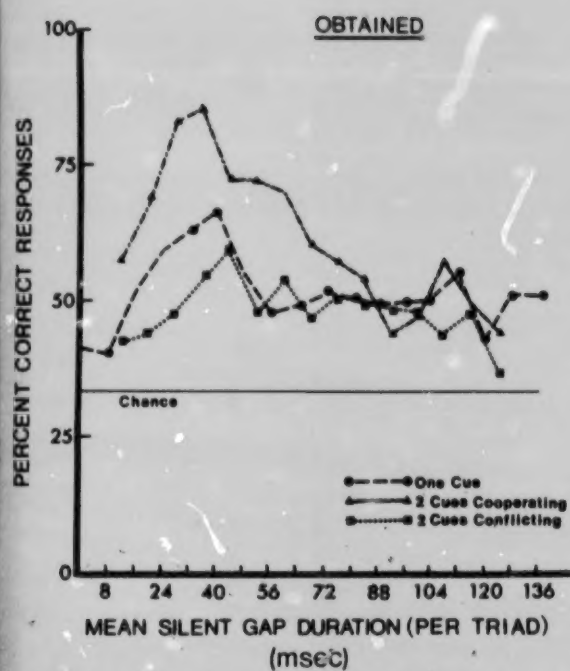


Figure 4. Obtained (left-hand panel) and Predicted (right-hand panel) results for the 3-way oddity test in Experiment 1.

Table 1

ω Values of Tukey tests on significant Comparison Type (CT) effects from ANOVAs for Experiments 1 and 2.

	ONE CUE vs. COOPERATING CUES	COOPERATING CUES vs. CONFLICTING CUES	CONFLICTING CUES vs. ONE CUE
<u>EXPERIMENT 1</u> (represented as mean percentages correct):			
Obtained data:	8.72++	13.16++	4.44*
Predicted data:	9.57+	18.14++	8.58+
Obtained "peak-range" data:	17.22++	25.97++	8.75*
Predicted "peak-range" data:	25.91++	43.22++	17.31+
<u>EXPERIMENT 2</u> (represented as mean true d' values):			
Obtained data:	1.43++	2.04++	.61*
Predicted data:	1.85++	2.13++	.29
Observed "peak-range" data:	1.68++	2.88++	1.20**
Predicted "peak-range" data:	2.31++	3.32++	1.01**
<u>EXPERIMENT 3</u> (represented as mean number correct, out of 6):			
"Say"- "stay" listeners:	.72+	1.08++	.36*
"Temporal" listeners (peak):	.76*	.91*	.15

* $p < .05$
 ** $p < .01$
 + $p < .005$
 ++ $p < .001$

Predicted only for comparisons distant from the between-category performance peak by 16-24 msec or more. These interactions also indicated that the Obtained peaks were shifted slightly toward the D boundary (highest performance levels in the "cooperating cues" and "one cue" comparisons at approximately 36-40 msec), relative to the Predicted peak positions, which were at the average of the crossover boundary values for the two continua (highest performance levels at approximately 44-48 msec).

Because differences between Obtained and Predicted performance were not found for between-category comparisons, the pattern of within-category differences was examined. For this purpose, within-category refers to all comparisons whose predicted level was 39% or less ($\leq 6\%$ above chance), since in each of those comparisons the stimuli being compared never differed by more than 10% in their category assignments in the identification test. This 'within-category' range includes "conflicting cues" comparisons with mean gap durations (per comparison pair) of between 12-44 msec and between 60-128 msec, "one cue" comparisons with gaps of between 0-16 msec and between 56-136 msec, and "cooperating cues" comparisons with mean gap durations of between 76-128 msec. The pattern of mean above-chance discrimination for these within-category comparisons was "cooperating cues" (mean correct = 49.2%) > "one cue" (48.2%) > "conflicting cues" (47.0%). If we consider only the within-category Obtained performance that was unaccounted for by phonetic category identifications (residual performance = Obtained mean minus Predicted mean), the pattern remains: "cooperating cues" (residual performance = 14.0%) > "one cue" (13.0%) > "conflicting cues" (11.7%). These small but consistent Obtained vs. Predicted differences will be discussed later in the paper, since the effects are best understood relative to the Obtained-Predicted differences found in Experiments 2 and 3.

Individual patterns. Consistency of the "phonetic" perceptual equivalence pattern in individuals was also examined, for stimulus comparisons with average gap durations near the two identification boundaries (between 20 - 48 msec), henceforth termed the "peak-range." This analysis was based on the mean percent correct Obtained responses for "one cue" comparisons between s[24]D-s[24]d and s[48]D-s[48]d, "cooperating cues" comparisons between s[32]D-s[8]d and s[56]D-s[32]d, and "conflicting cues" comparisons between s[32]d-s[8]D and s[56]d-s[32]D]. Ten of the 15 subjects showed the order "cooperating cues" > "one cue" > "conflicting cues," a proportion (67%) significantly better than chance by binomial test, $p < .002$. For the other five subjects, "one cue" performance was equal to or lower than "conflicting cues" performance. However, the key criterion for "phonetic" perceptual equivalence is that "peak-range" performance for "cooperating cues" be better than for "conflicting cues," since both involve contrast on two acoustic dimensions. This latter criterion was met by all 15 subjects, a proportion that is far beyond chance expectation by binomial test, $p < .001$. Two one-way ANOVAs performed on the Obtained and the Predicted "peak-range" means, for the two Comparison Types, supported the "phonetic" predictions even more strongly than did the overall group ANOVAs. For the Obtained "peak-range" data, the Comparison Types effect was significant, $F(2,28) = 36.59$, $p < .001$, as it was for the Predicted "peak-range" data, $F(2,28) = 81.51$, $p < .001$ (see Table 1 for all Tukey pairwise contrasts).

Conclusions

The results of Experiment 1 clearly indicated a trading relation and perceptual equivalence between gap duration and F1 onset frequency as cues for the "say"- "stay" distinction. The Fitch et al. (1980) findings were thus replicated for a different phonetic category contrast. Silent gap duration and F1 onset frequency appeared to have converged on a single dimension in phonetic perception.

EXPERIMENT 2

Although Experiment 1 suggested that the silence and F1 spectral cues for the "say"- "stay" contrast are perceptually equivalent in quality, the oddity task's heavy demands on auditory short-term memory⁶ may have biased the subjects to recode the stimulus properties into phonetic category information. Phonetic categorizations are believed to be better retained in memory than are raw acoustic properties, especially in the case of consonants (cf. Crowder, 1971, 1973; Crowder & Morton, 1969; Darwin & Baddeley, 1971; Fujisaki & Kawashima, 1968; Pisoni, 1971, 1975; Pisoni & Tash, 1974; Repp, Healy, & Crowder, 1979). Insofar as the oddity task encourages listeners to (phonetically) recode rapidly fading sensory information, it would not be a sensitive test for qualitative equivalence of cues at the sensory level. For this reason, Experiment 2 used a 2IAX ("same"- "different") discrimination procedure to induce performance reflecting greater perceptual sensitivity to the physical properties of the stimuli. To further ease memory demands, short ISIs were used. In addition, the identification test was run after the discrimination task, to control against the possibility that obtaining forced-choice identifications before discrimination judgments might also have introduced an experimental bias toward phonetic categorization (and thus against comparison of physical properties).

We used the 2IAX procedure to assess perceptual sensitivity, based on several considerations.⁷ Whereas standard Signal Detection Theory (SDT-- MacMillan, Kaplan, & Creelman, 1977) cannot be applied to estimation of perceptual sensitivity from oddity data, the SDT framework does permit perceptual sensitivity to be estimated from ABX, 4IAX, and 2IAX data, through the use of the d' sensitivity index. The ABX task should be the most sensitive of those three paradigms, according to SDT predictions (MacMillan et al., 1977) that d' values should be highest for ABX and lowest for 2IAX; however, the speech and nonspeech data on relative sensitivities of the paradigms are equivocal. Higher d' values have been obtained through ABX than 4IAX tests for frequency (Creelman & MacMillan, 1979) and intensity discriminations (Pastore, Friedman, & Baffuto, 1976), but 4IAX tasks have obtained higher d' values than ABX for phase discriminations (Creelman & MacMillan, 1979) and speech discriminations (with a larger task effect for vowels than consonants--e.g., Pisoni, 1971, 1975; Pisoni & Lazarus, 1974). Two 4IAX observer strategies are logically possible, and one is much more sensitive than the other; inability to predict which strategy listeners will adopt (Creelman & MacMillan, 1979; MacMillan et al., 1977) made a choice between ABX

and 4IAX for Experiment 2 difficult. Moreover, a recently-published formula for computing bias-corrected d' values from 2IAX data (true d') yields d' values at least as high for 2IAX data as for AXB and 4IAX (Kaplan, MacMillan, & Creelman, 1978).⁸ Finally, memory demands for making "same"- "different" judgments on two "say"- "stay" stimuli per trial seemed to us to be lower than for AXB or 4IAX judgments.

Method

Subjects

A new group of 14 subjects participated in this experiment. Seven were Yale undergraduates; the other seven were undergraduates at the University of Massachusetts/Amherst. All reported having normal hearing in both ears.

Stimuli

The stimuli from Experiment 1 were used again. This time, however, stimuli containing gaps of over 96 msec were eliminated from the test, since they had been identified as "stay" nearly 100% of the time in Experiment 1 and were not crucial to either the identification or the discrimination test. The truncated "say"- "stay" continua contained 14 stimuli each.

Procedure

The subjects first completed a 3-way 2IAX test, which employed 300 msec ISIs and 2.5 sec ITIs. Stimulus pairs for the three types of test comparisons ("one cue," "cooperating cues," and "conflicting cues") were chosen from the truncated "say"- "stay" continua by the same means as described in Experiment 1, with 24 msec silence again used as the temporal compensation value in both two-cue comparison types. In addition, a fourth set of "physically same" catch-trial comparisons was included to provide false alarm rate data. There were 11 possible mean gap values each for "one cue" ($s[8]d-s[8]D$ to $s[88]d-s[88]D$) and "physically same" comparisons ($s[8]d-s[8]d$ and $s[8]D-s[8]D$, to $s[88]d-s[88]d$ and $s[88]D-s[88]D$), and 10 possible pairs each for "conflicting cues" (from $s[0]D-s[24]d$ to $s[72]D-s[96]d$) and "cooperating cues" comparisons (from $s[0]d-s[24]D$ to $s[72]d-s[96]D$).

Six judgments were obtained for each of the 42 possible stimulus contrasts (total items = 252). The test sequence was randomized across all stimulus pairings and comparison types. To help focus attention on the differences in acoustic properties of the stimuli, rather than on phonetic categories, the subjects were told that most AX pairs were physically different. They were told that most of the "different" pairs would be tokens of the same word, so they should listen closely for slight sound differences between members of same-word pairs.

A randomized 280-trial forced-choice identification test (2.5 msec ISIs), containing 10 repetitions of all stimuli in the truncated continua, was administered after the 2IAX test.

Results and Discussion

Identification Test

The identification results (lower panel, Figure 5) replicated the trading relation found in Experiment 1, this time with a boundary difference of 18.5 msec (range = 5.1-37.3 msec). The somewhat smaller trading relation may have resulted from truncating the continua (a possible response bias, or stimulus range, effect), but was nonetheless significant, $t = 8.88$, $p < .001$. The D (strong "day") category boundary fell at a smaller gap duration (mean = 25.3 msec, range = 10.7-51.2) than the d (weak "day") boundary (mean = 43.8, range = 33.6-60.0) for all individuals.

Three-way 2IAX Test

The 3-way 2IAX results (upper panel, Figure 5) showed the "phonetic" order of "different" response levels for the three types of test comparisons, as did the 3-way oddity results of Experiment 1. The small "peak" in percentage of "different" responses for "physically same" catch-trials near the category boundary results from the ambiguous identifications of the boundary stimuli.

The 2IAX results suggest that the two cues were perceptually equivalent; however, d' values offer a more accurate measure of differential perceptual sensitivities among the three test comparisons than do the raw data, since the 2IAX task biases observers to give "same" responses for physically-different pairs that are difficult to discriminate. Although the d' values computed by the standard formula [$z(\text{Hits}) - z(\text{False Alarms})$] are artificially deflated by this bias, a revised formula for bias-corrected d' values estimated from 2IAX data, and the resulting table of true d' values, have recently been published (Kaplan et al., 1978). Therefore, we determined true d' values for the Obtained 2IAX data, and for the Predicted data computed from the identification results (formulae in Appendix C).

Figure 6 shows the Obtained (left-hand panel) and Predicted sensitivity functions (right-hand panel) represented by the true d' values. Several ANOVAs were run on the true d' data, spanning the range of overlap among the three comparison types (12-88 msec mean gaps). For both the Obtained and the Predicted data, two-way ANOVAs were performed for three Comparison Types x 10 Stimulus Pairs. In addition, for each of the three comparison types, ANOVAs were performed for two Functions (Obtained vs. Predicted) x 10 Stimulus Pairs. Finally, the mean d' values for both the Obtained "peak-range" data and the Predicted "peak-range" data (20-48 msec mean gaps--cf. Experiment 1) were entered into one-way ANOVAs on the three Comparison Types. These ANOVAs were analogous in design to those run on the Experiment 1 data, and the pattern of results (see Table 2, and Tukey pairwise contrasts in Table 1) essentially replicated the findings from that previous experiment.

The results deviated slightly from the Experiment 1 pattern--the 2IAX test, in contrast to the oddity test, produced a small but significant performance-level peak (higher d' values) near the category boundary for both

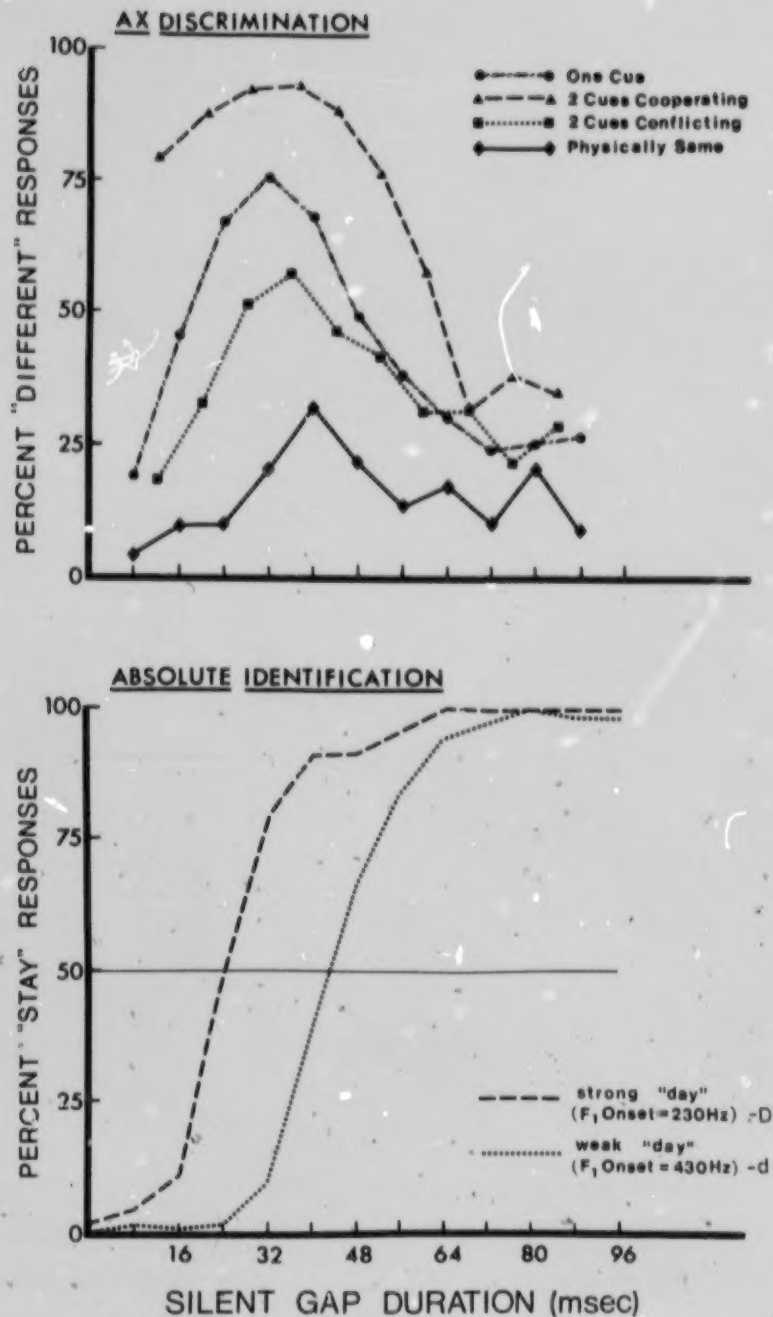
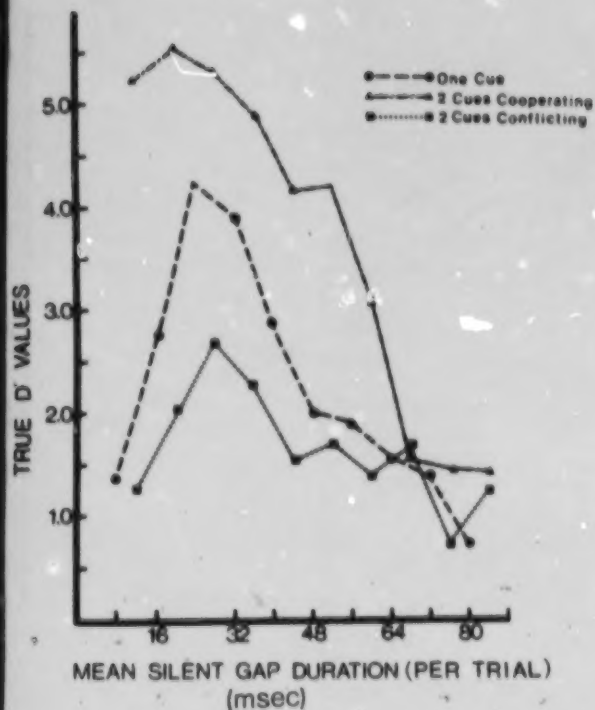


Figure 5. Obtained functions for the 3-way 2IAX discrimination test (upper panel) and the forced-choice identification test (lower panel), for Experiment 2.

OBTAINED



PREDICTED

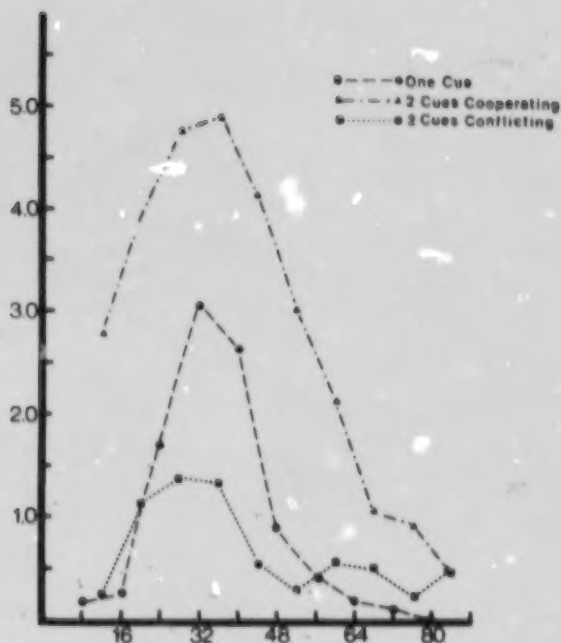


Figure 6. Obtained (left-hand panel) and Predicted (right-hand panel) functions for true d' values on the 3-way 2IAX test in Experiment 2.

Table 2

Results from the ANOVAs performed on the true d' values computed for the 3-way 2IAX data from Experiment 2.

ANOVA	Effect	df	F	p
<u>Obtained data</u> (3CTa x 10SPb):	CT	2,26	41.78	<<.001
	SP	9,117	17.08	<<.001
	CT x SP	18,234	6.48	< .001
<u>CT x SP Simple Effects:</u>				
"cooperating cues"	SP	9,351	18.43	<<.001
"one cue"	SP	9,351	8.47	< .001
"conflicting cues"	SP	9,351	1.97	< .05
<u>Predicted data</u> (3CT x 10SP):	CT	2,26	109.99	<<.001
	SP	9,117	17.88	<<.001
	CT x SP	18,234	9.42	< .001
<u>CT x SP Simple Effects:</u>				
"cooperating cues"	SP	9,351	26.91	<<.001
"one cue"	SP	9,351	13.39	< .001
"conflicting cues"	SP	9,351	1.83	> .05c
<u>"Cooperating cues"</u> (2Fd x 10SP):	F	1,13	17.26	<<.001
	SP	9,117	37.96	<<.001
	F x SP	9,117	2.09	< .05
<u>"One cue"</u> (2F x 10SP):	F	1,13	35.23	<<.001
	SP	9,117	13.02	<<.001
	F x SP	9,117	1.90	> .05e
<u>"Conflicting cues"</u> (2F x 10SP):	F	1,13	21.11	<<.001
	SP	9,117	4.58	< .001
	F x SP	9,117	.49	n.s.
<u>Obtained "peak-range" data</u> f (3CT):	CT	2,26	31.66	<<.001
<u>Predicted "peak-range" data</u> (3CT):	CT	2,26	66.23	<<.001

a Comparison Types ("cooperating cues", "one cue", "conflicting cues")

b Stimulus Pairs

c marginal (.05 cut-off = 1.88)

d Functions (predicted vs. observed)

e marginal (.05 cut-off = 1.96)

f mean value for gap durations between 20-48 msec (average per AX pair) in each type of test comparison

the Obtained and Predicted "conflicting cues" comparisons. However, this deviation is not damaging to the claim that Experiment 2 replicated the results of Experiment 1. The peak reflects the fact that the trading relation found in this experiment was only 18.5 msec, which did not match the predetermined 24 msec gap compensation value used. Since gap duration and F1 onset differences in the "conflicting cues" comparisons did not precisely cancel one another, there was a small, but predictable, enhancement of sensitivity near the boundary.

Obtained vs. Predicted differences. The pattern of differences between Obtained and Predicted functions was virtually the same as in Experiment 1 (see Table 2, "Cooperating cues," "One cue," and "Conflicting cues"), demonstrating a higher sensitivity for Obtained than for Predicted data, particularly for comparisons beyond the immediate range of the between-category performance peaks. Mean residual sensitivities were calculated (Obtained true d' minus Predicted true d') for the three comparison types, across all comparisons in which the pair members had deviated from one another by less than 10% in proportion of "stay" identifications (roughly 'within-category' pairs, as defined in Experiment 1); included were "conflicting cues" comparisons with mean gap durations (per pair) of 12 msec and between 44-84 msec, "one cue" comparisons with gaps of 8 msec and between 48-88 msec, and "cooperating cues" comparisons with mean gaps of between 60-84 msec. These Obtained-Predicted difference values followed the order "cooperating cues" (mean residual sensitivity = 1.49) $\geq$ "one cue" (1.46) $>$ "conflicting cues" (1.37). However, because of the shift in positions of the Obtained peaks relative to the Predicted peaks, we recalculated residual sensitivity values (Obtained-Predicted differences) according to deviations in gap duration from the value at the peak. The order of these residual sensitivities (Figure 7) was "cooperating cues" (1.58) $>$ "one cue" (1.38) $>$ "conflicting cues" (.98). Once again, detailed discussion of Obtained vs. Predicted differences will be deferred until later in the paper.

Individual patterns. The individual patterns of mean "peak-range" sensitivities were virtually identical to those found in Experiment 1. Nine of the 14 subjects, $p < .001$ by binomial test, showed the expected order: "cooperating cues" $>$ "one cue" $>$ "conflicting cues." However, the key criterion ("cooperating cues" $>$ "conflicting cues") was met by all 14 subjects, $p < .0002$.

Conclusions

All of the major findings from Experiment 1 were replicated--a trading relation was found between the F1-onset and silence cues, and the pattern of 2IAX sensitivities fit the "phonetic" predictions. The two acoustic cues for the "say"- "stay" contrast appear to be perceptually equivalent, even under conditions designed to reduce demands on auditory short-term memory and to reduce experimentally-induced biases to categorize stimuli before discriminating them. The results suggest that the equivalence of the silence and F1 onset frequency contrasts may reflect listeners' perceptual sensitivity to those particular physical properties of the synthetic "say"- "stay" stimuli.

RESIDUAL SENSITIVITY

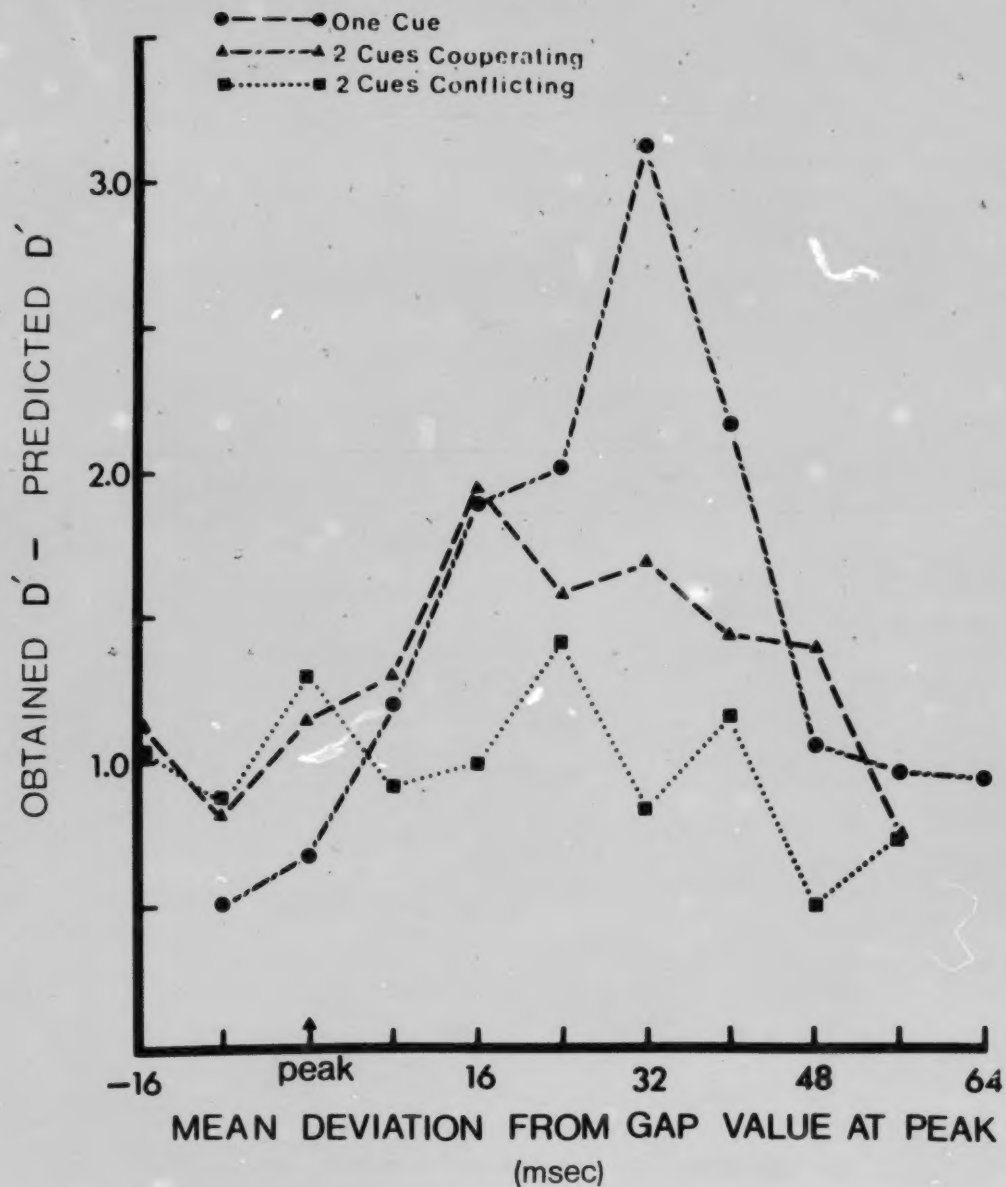


Figure 7. Residual sensitivity (Obtained d' minus Predicted d') found in the 3-way 2IAX test, Experiment 2.

EXPERIMENT 3

The question now arises, however, What is the origin of the equivalence in perceptual sensitivity to those two phonetic cues? At least two possibilities present themselves: (1) the "phonetic" alternative, that the trading relation/perceptual equivalence pattern derives specifically from perception of phonetic information (recall that even in Experiment 2, the subjects perceived the stimuli as "say" and "stay")--that is, it occurs "only for sounds ... being processed as speech" (Fitch et al., 1980, p. 344); or (2) the "psychoacoustic" alternative, that the pattern derives from general (not speech-specific) properties of auditory perception. Although we know of no research on psychoacoustic integration of acoustic cues like those we used in the "say"- "stay" research, the "psychoacoustic" alternative gains converging support from: (a) known trade-offs in nonspeech perception (e.g., the time-intensity trade in auditory localization: Green, 1976); and (b) speech-relevant discontinuities in perceptual responses to change in nonspeech stimuli along a single acoustic dimension (e.g., categorical nonspeech perception for rise-time and onset-time contrasts: Cutting & Rosner, 1974, 1976; Cutting, Rosner, & Foard, 1976; Miller, Wier, Pastore, Kelly, & Dooling, 1976; Pastore, Ahroon, Baffuto, Friedman, Puleo, & Fink, 1977; Pisoni, 1977).

Therefore, it was important to determine whether there was some psychoacoustic interaction between the "say"- "stay" cues. For example, it could be that longer gaps were needed for the d stimuli than for the D stimuli to be heard as "stay," because the /s/ offset was closer in frequency to the 430 Hz F1 onset than to the 230 Hz F1 onset (by 3-4 critical bands). This possibility seems unlikely, though, because it contradicts findings that gap sensitivity is inversely related to the amount of frequency difference between the acoustic components surrounding the gap (Divenyi, 1979; Divenyi & Danner, 1977; Divenyi & Sachs, 1978). Hence, the cue integration we found in Experiments 1 and 2 (and that found by Fitch et al., 1980) may indeed be unique to the perception of phonetic information. This last conjecture would not be at odds with the results of Experiment 2, since those subjects perceived the synthetic speech stimuli as "say" and "stay" (i.e., the equivalence in sensitivity to the silence and F1-onset cues might still have been dependent on perception of phonetic information). Nonetheless, the possibility remains open that the inverse relationship between temporal and spectral sensitivity might be reversed, for purely (and as yet unknown) psychoacoustic reasons, under certain stimulus and task conditions like those used in the "say"- "stay" tests. A third experiment was run to determine whether the "phonetic" or the "psychoacoustic" alternative would better explain trading relations and perceptual equivalence between phonetic cues.

This test required nonspeech control stimuli that, on one hand, maintained the crucial temporal and spectral properties of the "say"- "stay" stimuli, since the potential psychoacoustic effects might be dependent upon that particular array of physical properties. On the other hand, however, the stimuli had to be dissimilar enough from the "say"- "stay" stimuli that most naive listeners would fail to hear them as speech. We used 'sinewave analogues' of the "say"- "stay" continua for this experiment because they fit both criteria--they were essentially identical to the synthetic speech stimuli, except that their "formants" had bandwidths of 1 Hz. Another recent study (Bailey, Summerfield, & Dorman, 1978; Dorman, 1979) employed sinewave

analogues of speech continua, and reported that most naive listeners heard the sinewave stimuli as nonspeech sounds (e.g., beeps, chimes, slide-guitar notes, electronic tones). Only a few listeners spontaneously perceived them as distorted ('chime-like') speech. Identification tests revealed distinct differences in category boundaries, dependent on whether the sinewave stimuli were perceived as speech or as nonspeech. Because sinewave analogues can be perceived either as speech or as nonspeech, and because the perceptual mode can affect identification performance, we grouped our subjects according to their posttest reports of what the stimuli sounded like. The "phonetic" alternative predicted that the trading relation and perceptual equivalence between cues would occur only for subjects who heard the sinewave analogues as "say" and "stay." In contradistinction, the "psychoacoustic" alternative predicted that those perceptual patterns would occur even for subjects who perceived the sinewave stimuli as nonspeech.

Method

Subjects

Twenty-two naive listeners completed this experiment. Fifteen were Yale undergraduates, and seven were enrolled at the University of Massachusetts/Amherst. All reported normal hearing in both ears.

Stimuli

Sinewave analogues of the weak "day" (d) and strong "day" (D) speech syllables were made by synthesizing three simultaneous sinewaves, using a software program developed for the PDP-11/45 at Haskins Laboratories.⁹ In each of the sinewave "day" analogues, the time-varying amplitude and center-frequency characteristics of each of the synthetic speech formants was imitated by a frequency- and amplitude-modulated sinewave (see Figure 8). The sinewave analogue of D will be termed 'strong SW1 transition' (SW), and the analogue of d will be termed 'weak SW1 transition' (sw).¹⁰

The nonspeech analogue of /s/ also had to differ enough from natural /s/ to be heard as a nonspeech sound by most listeners, and yet be similar enough to /s/ that it could be heard as speech. We used a "hiss" created by changing the rise-time and overall amplitude (but not the fall-time) of the natural /s/ used in the synthetic "say"- "stay" continua (see Figure 9). The hiss (abbreviated 'h') had the same frequency and offset properties as the /s/ of Experiments 1 and 2, and thus met the requirements just outlined.

The SW and sw continua were constructed by inserting varying gap durations (in 8 msec steps, from 0-96 msec, as in Experiment 2) between h and the sinewave analogue. Each continuum contained 14 stimuli (h[0]SW to h[96]SW, and h[0]sw to h[96]sw), which most of our naive listeners heard as bizarre electronic sounds or distorted nonspeech sounds, such as beeps, water drips, etc. Only about one-quarter of the subjects perceived them as 'chime-like' utterances of "say" and "stay."

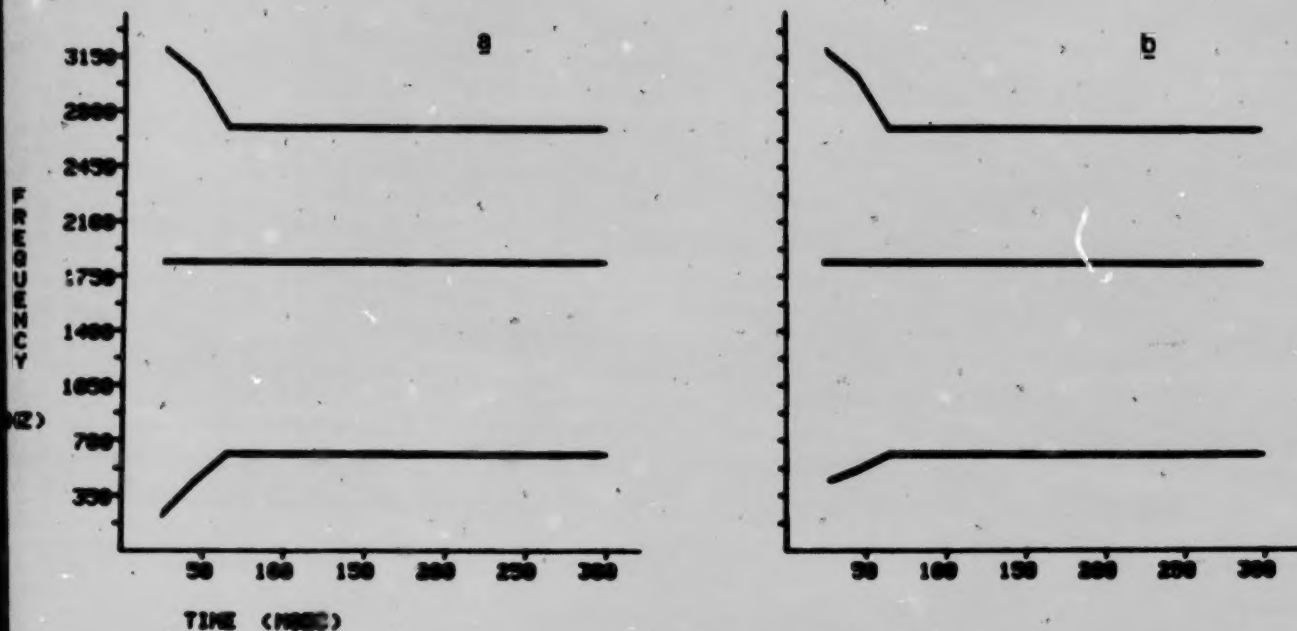


Figure 8. Schematic time-spectrum representations of the 3-sinewave analogues for the synthetic speech syllables used in Experiments 1 and 2: (a) weak "day" analogue, (b) strong "day" analogue.

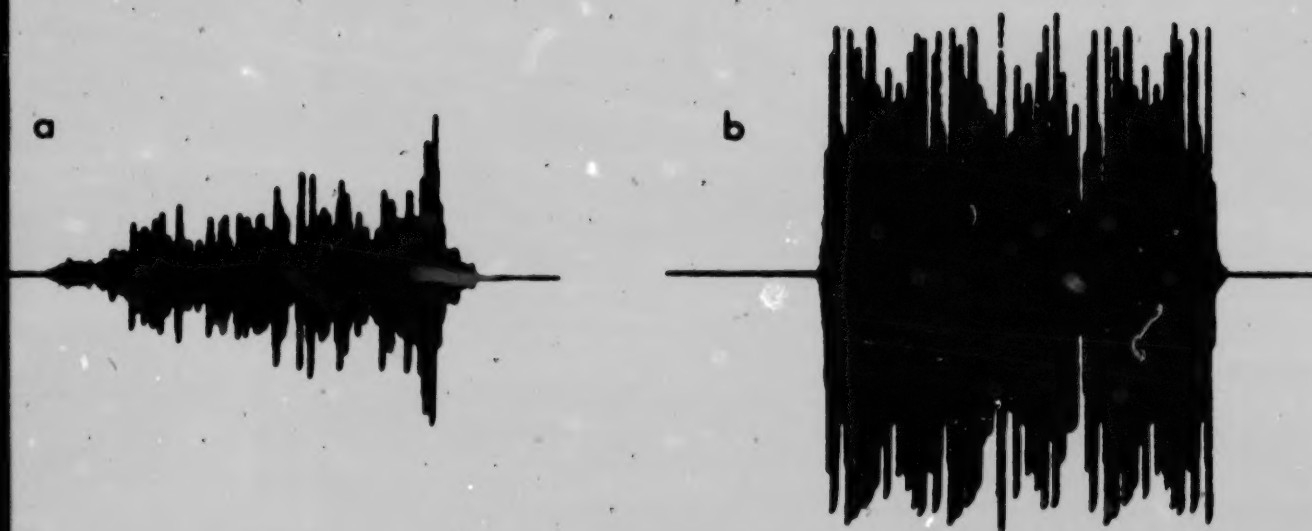


Figure 9. CRT display of digitized waveforms for (a) natural male /s/ (from "stay") used in the "say"- "stay" continua for Experiments 1 and 2, and (b) amplitude-resaped 'hiss' used in the sinewave continua for Experiment 3.

Procedure

An AXB identification procedure was used for collecting categorization judgments (cf. Bailey et al., 1978; Dorman, 1979), since a standard forced-choice identification test would have required category labels, which might have encouraged subjects to perceive the stimuli as speech. On each trial, the second stimulus (X) had to be identified as being more similar either to the first stimulus (A) or the last stimulus (B). Categories 'A' and 'B' were fixed across trials; A was the closest sinewave analogue to natural "say" (h[0]sw, the analogue for s[0]d), and B was the closest analogue to natural "stay" (h[96]SW, the analogue for s[96]D). Each of the 28 possible AXB trials was presented 10 times in a randomized test sequence.

After the AXB test and a subsequent 15-minute break, the subjects took a 3-way oddity test, which was designed exactly as in Experiment 1. It included "cooperating cues" (10 contrasts: h[0]sw-h[24]SW to h[72]sw-h[96]SW), "one cue" (13 contrasts: h[0]sw-h[0]SW to h[96]sw-h[96]SW), and "conflicting cues" comparisons (10 contrasts: h[0]SW-h[24]sw to h[72]SW-h[96]sw). Six judgments were obtained for each of the 33 possible stimulus contrasts, and the test sequence was randomized across all comparison types.¹¹

Group assignments. Sixteen subjects (all 7 University of Massachusetts subjects and 9 of the Yale subjects) were told before testing that the stimuli were computer sounds with two components--a 'hiss' followed by a 'chime-like' sound, with varying gap lengths between the components. The remaining six Yale subjects were told that the stimuli were distortions of "say" and "stay," and that they should listen for those words as they completed their tasks. We hoped this would induce a "speech perceptual set," and thereby allow us to assess the contribution of speech processing/perception to performance.

Subjects answered a posttest questionnaire on what the stimuli sounded like, and which stimulus properties they had attended to. The subjects were divided into five groups, according to their questionnaire responses. One subgroup included four subjects who claimed to have been guessing or changing their perceptual strategies from trial to trial; their performance was near chance, and appeared haphazard. Another group of three subjects perceived speech contrasts other than "say"- "stay" (i.e., "sleh"- "sreh," the French "un"- "rien," and "two Greek words"), including one subject who had been instructed to listen for "say" and "stay." The perceptual patterns for these first two groups will not be discussed further. Only the remaining three subgroups will be discussed in more detail:

- (1) Five subjects who heard "say" and "stay," either by instruction (3 subjects) or spontaneously (2 subjects), for even a portion of the test session. Three claimed to have occasionally listened for tone differences (the 2 'spontaneous' subjects) or for different water drips (1 'instructed' subject, who used this strategy throughout most of the 3-way oddity test), because at times they 'lost touch with' the words. Therefore, this grouping provides a conservative test of the "phonetic" alternative.
- (2) Five subjects who focused primarily on nonspeech temporal contrasts related to changes in gap duration, including differences in the gaps

(2 subjects) or spaces (1 subject) between the hiss and sinewaves, and overall differences (2 subjects). (Note that overall duration was confounded with gap duration in all three experiments of this study.)

- (3) Five subjects who perceived nonspeech contrasts related to the SW vs. sw spectral difference. These subjects generally ignored the hiss, and listened for contrasts between two different kinds of water drips (2 subjects), two different pitches (1 subject), the presence or absence of a ringing quality (1 subject), or the presence or absence of electronic "waw" (1 subject).

Groups 2 and 3 each included one of the subjects instructed to listen for "say" and "stay"; each reported that they could not hear the words, and had instead perceived a nonspeech contrast.¹² The three groups of subjects will be referred to as: (1) "say"- "stay" listeners, (2) "temporal" listeners, and (3) "spectral" listeners.

Results and Discussion

AXB Identification Test

Group comparisons. The AXB identification data for the three groups are shown in the upper panels of Figures 10-12, with percentage of 'Category B' responses (h[96]SW) on the ordinate. The figures suggest fairly large differences in the way the groups categorized the stimuli in the two sinewave continua. To determine whether the AXB differences among the three groups were statistically significant, a two-way ANOVA was performed for three Groups x two Continua. The data used in this test were "percent Category B responses" because the majority of AXB functions for the "spectral" group had no 50% crossovers within the range tested. The significant Groups x Continua interaction, $F(2,12) = 14.77$, $p < .01$, indicates that categorizations of the SW and sw continua contrasted substantially among the three groups of listeners. Only the "say"- "stay" group showed the sort of trading relation found in Experiments 1 and 2.

"Say"- "stay" listeners. The "say"- "stay" listeners showed a 17.8 msec trading relation (range = 7.6-29.6) between gap duration and the SW-sw spectral contrast (upper panel, Figure 10). The boundary difference was significant by a one-way ANOVA, $F(1,4) = 28.8$, $p < .01$, and was quite close in magnitude to that found for the corresponding speech patterns in Experiment 2, which included the same range of gap durations. All five subjects showed larger crossover values for the sw continuum (mean = 45.8 msec, range = 40.0-52.0) than the SW continuum (mean = 27.9 msec, range = 22.4-38.4); these category boundaries were nearly the same as in Experiment 2.

"Temporal" listeners. In contrast to the "say"- "stay" group, the "temporal" listeners (upper panel, Figure 11) failed to use the SW-sw contrasts consistently in their categorizations. They categorized stimuli according to gap duration (or overall duration), showing a SW boundary at 34.3 msec (range = 26.0-44.8), and a sw boundary at 40.7 msec (range = 32.0-48.0). However, the 6.4 msec boundary difference (range = -2.65 to +8.0) was not

"SAY-STAY" LISTENERS

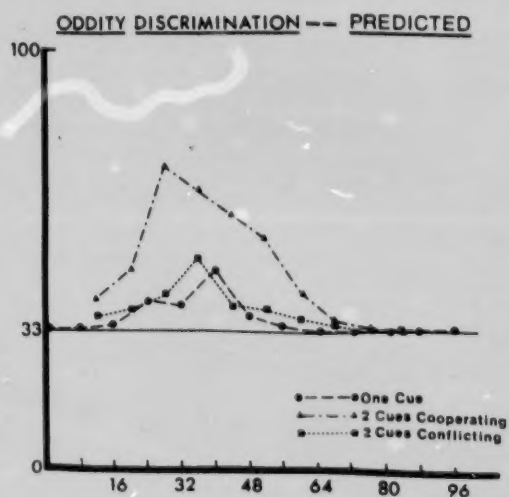
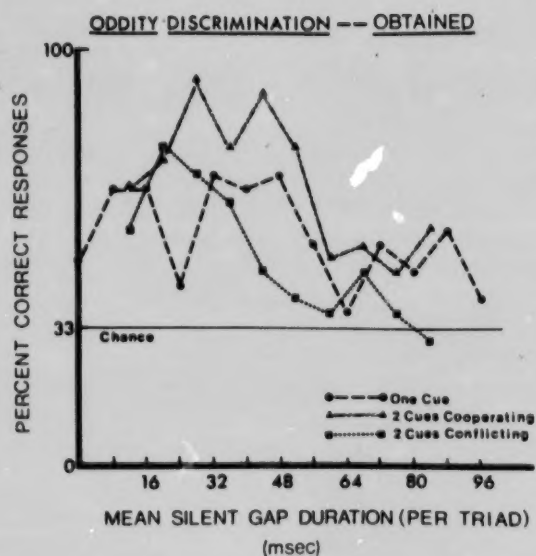
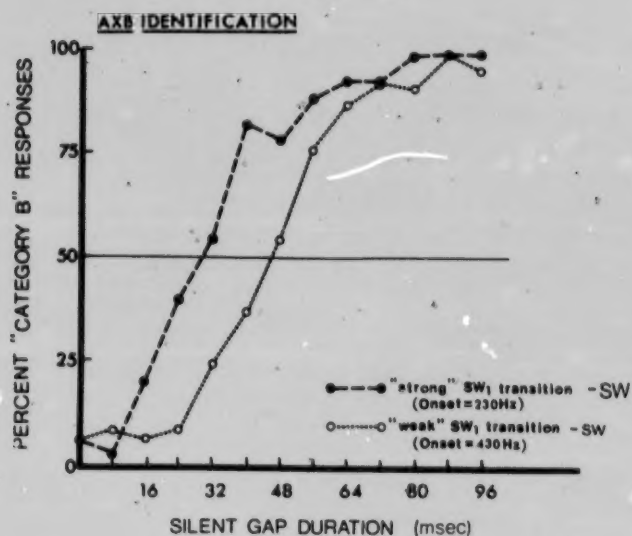


Figure 10. Sinewave AXB identification functions (upper panel), and Obtained (lower left) and Predicted (lower right) functions for the 3-way oddity test, "say"- "stay" listeners, Experiment 3.

"TEMPORAL" LISTENERS

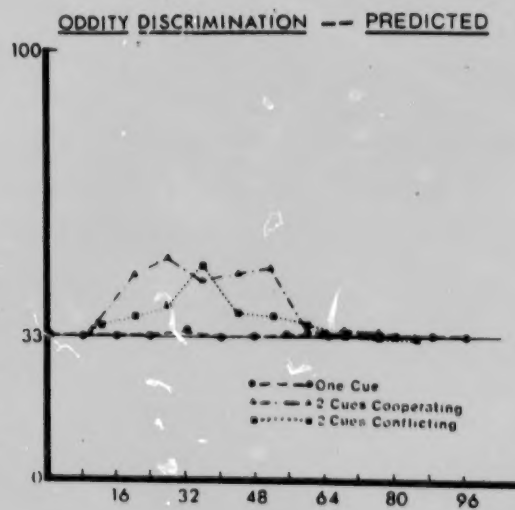
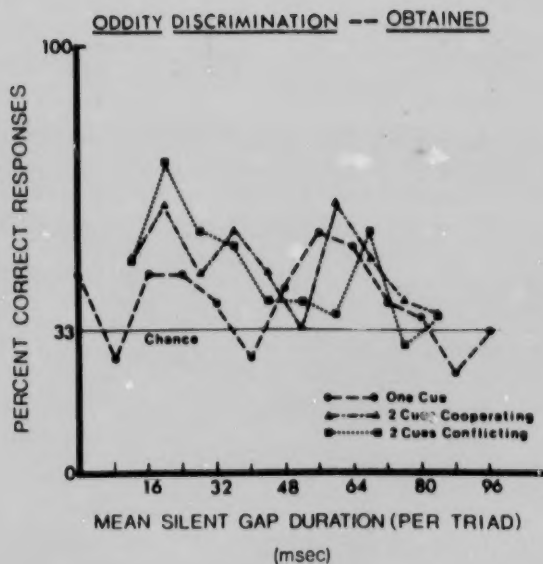
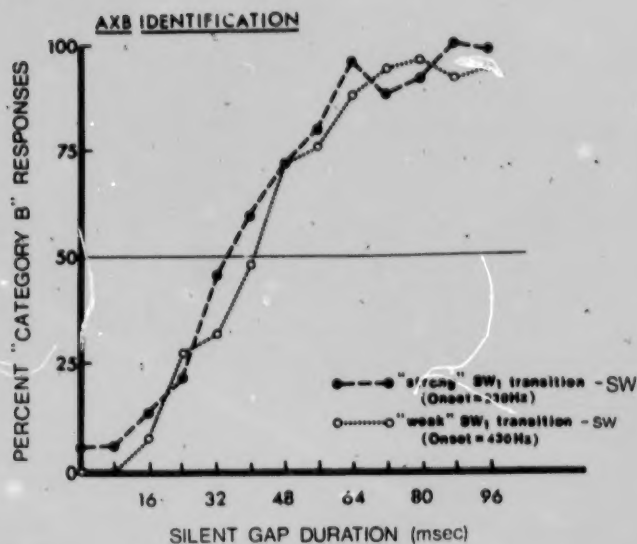


Figure 11. Sinewave AXB identification functions (upper panel), and Obtained (lower left) and Predicted (lower right) functions for the 3-way oddity test, "temporal" listeners, Experiment 3.

"SPECTRAL" LISTENERS

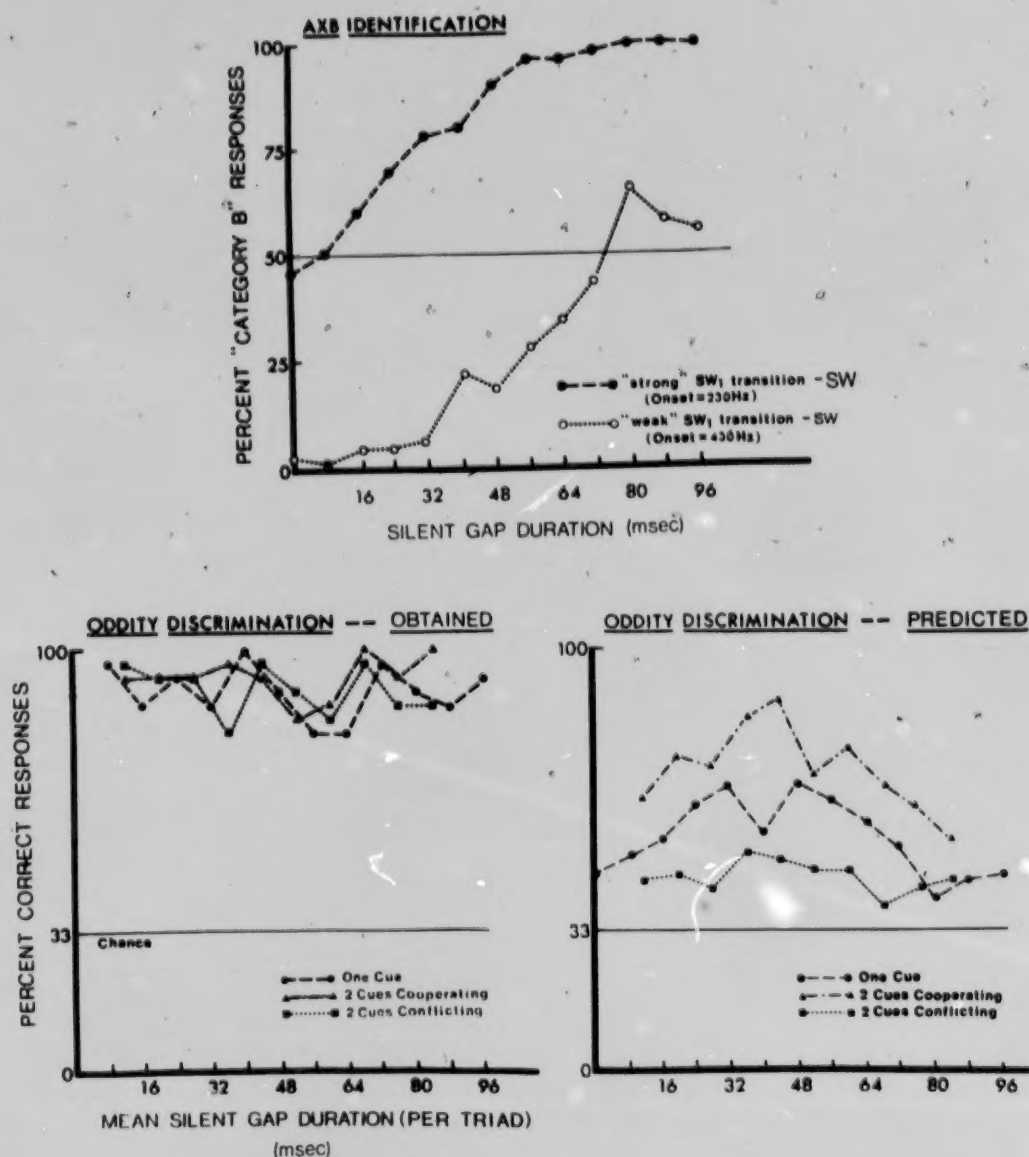


Figure 12. Sinewave AXB identification functions (upper panel), and Obtained (lower left) and Predicted (lower right) functions for the 3-way oddity test, "spectral" listeners, Experiment 3.

significant, according to their simple effects test for the Groups x Comparison Types interaction.

"Spectral" listeners. In contrast to the other two sinewave groups, the "spectral" listeners consistently categorized stimuli according to SW-sw differences (upper panel, Figure 12), according to their simple effects test for the Groups x Continua interaction, $F(1,12) = 60.1$, $p < .001$. None of the SW stimuli were identified with A more often than chance, nor were the SW stimuli identified with B more often than chance, except for one token categorized as B 65% of the time. Thus, lengthening the gaps did not completely compensate for the SW-sw difference--for the longest gaps, the SW function turned back toward the 50% line rather than toward higher percentages of B responses. The asymptote of the SW function at long gaps makes it unlikely that extending the range of the gap durations would have resulted in a complete trading relation for this group. Thus, the "spectral" listeners failed to show the complete trading relation found with perception of "say"- "stay," and they failed in a qualitatively different way from the "temporal" group.

Three-way Oddity Test

Group comparisons. There were also group differences in the pattern of 3-way oddity discrimination (Figures 10-12) for both the Obtained (lower left panels) and the Predicted data (lower right panels). To compare discrimination performance among the three groups, mean "peak-range" (16-48 msec gaps) performances were calculated for the three comparison types. The range was extended beyond the 20-48 msec "peak-range" used in Experiments 1 and 2 in order to include all the peaks of the "say"- "stay" group (lower left, Figure 10) and the peaks of the "temporal" group at 20 msec (lower left, Figure 11). An ANOVA was performed on these data for three Groups x three Comparison Types. The order of "peak-range" performance levels differed significantly among the three groups, Groups x Comparison Types: $F(4,24) = 3.49$, $p < .025$.

"Say"- "stay" listeners. Only the "say"- "stay" group (lower left, Figure 10) showed the "phonetic" pattern found in the previous two experiments (simple effects test for the Groups x Comparison Types interaction, $F(2,24) = 9.4$, $p < .01$). An ANOVA was run on their Obtained data for three Comparison Types x 10 Stimulus Pairs (mean gap durations of between 12-88 msec). The significant Comparison Types effect, $F(2,8) = 33.35$, $p < .001$, indicated that the order of performance was "cooperating cues" > "one cue" > "conflicting cues" (see Tukey pairwise contrasts, Table 1).¹³ Peak-level performance was significantly higher than trough-level performance across the three comparison types, according to the Stimulus Pairs effect, $F(9,36) = 4.36$, $p < .001$. Four of the five listeners in this group met the crucial criterion for "phonetic" perceptual equivalence ("cooperating cues" > "conflicting cues"), $p < .04$ by binomial test. The subject who failed had listened for "different water drips" during most of the discrimination test; her data show much the same pattern as the "spectral" group (described below). All four listeners who had more consistently perceived "say" and "stay" met the criterion, $p < .01$.

"Temporal" listeners. The Obtained discrimination pattern for this group (lower left, Figure 11) suggested that the "cooperating cues"-"conflicting cues" distinction was moot for them--what mattered were noticeable temporal differences between stimuli. The three Comparison Types x 10 Stimulus Pairs ANOVA performed on their Obtained data yielded a nonsignificant Comparison Types effect, which is not surprising since two-cue performance appeared to be better than "one cue" for less than one-third of the range (around the peak at 20 msec). However, the significant Stimulus Pairs effect, $F(9,36) = 3.08$, $p < .01$; indicates that there were two performance peaks; the level of the 20 msec peak appears higher than the one around 60-68 msec. The order of performance in the 20 msec "peak-range" (simple effects test for Groups x Comparison Types interaction, $F(2,24) = 4.86$, $p < .025$) was "conflicting cues" > "cooperating cues" > "one cue" (see Tukey pairwise contrasts, Table 1). That order fits the "auditory" prediction that performance for two-cue comparisons would be better than for "one cue."

"Spectral" listeners. The "spectral" listeners, unlike either of the other two sinewave groups, discriminated the SW-sw contrast nearly perfectly across the board (lower left, Figure 12). The nonsignificant effects in the three Comparison Types x 10 Stimulus Pairs ANOVA on their Obtained data indicated that performance was equally high across all three comparison types and at all gap durations, which was corroborated by the nonsignificant simple effects test for the Groups x Comparison Types interaction. All five subjects showed nearly perfect "peak-range" performance on all three comparison types.

Obtained vs. Predicted Differences

Group comparisons. All three groups of listeners showed better Obtained than Predicted discrimination (compare lower left panels and the corresponding lower right panels, Figures 10-12). However, the pattern of residual performance differed among the groups. In order to directly compare Obtained-Predicted differences among the groups, mean percentages of performance unexplained by phonetic categorizations (Obtained mean minus Predicted mean) were calculated for the entire range, within each comparison type.

"Temporal" listeners. The residual performance levels for the "temporal" group were small, and followed a different pattern from that observed in the two previous "say"-"stay" experiments: "conflicting cues" (mean residual performance = 8.4%) > "cooperating cues" (7.0%) > "one cue" (6.3%).¹⁴ Thus, their residual performance pattern, like their "peak-range" Obtained performance pattern, followed the "auditory" prediction of two-cue performance being better than "one cue." The most important contrast with the pattern found in Experiments 1 and 2 is that here the largest difference was "conflicting cues" > "one cue," which is crucial for the "auditory" hypothesis that two-cue performance should be better than "one cue" (and opposes the "phonetic" prediction that "one cue," as well as "cooperating cues," performance should be better than "conflicting cues").

"Spectral" listeners. The "spectral" group showed the highest residual performance levels observed thus far. Predicted performance followed the "phonetic" pattern, except that there were no peaks in performance. However, Obtained performance was near ceiling across the three comparison types.

Residual performance was very high, and showed the order: "conflicting cues" (mean residual performance = 43.9%) > "one cue" (35.2%) > "cooperating cues" (22.3%). Again, we see the "auditory" pattern of "conflicting cues" > "one cue," although the residual performance pattern was constrained by near-ceiling Obtained performance for all three comparison types. This residual performance pattern for the "spectral" listeners was very different from the "temporal" listeners, and from the "say"- "stay" listeners in the first two experiments.

"Say"- "stay" listeners. The residual performance pattern for the sinewave "say"- "stay" group was like the patterns found in the two previous "say"- "stay" experiments: "cooperating cues" (mean residual performance = 19.5%) > "one cue" (19.4%) > "conflicting cues" (14.6%). Also, the positions of the Obtained peaks were slightly shifted toward the SW category boundary, relative to the positions of the Predicted peaks, just as the Obtained peaks in the first two experiments had shifted toward the D boundary.

Even though the residual performance differences among comparison types were small in this group and in the first two experiments, the pattern was consistent, and was distinctly different from the "temporal" and "spectral" patterns. In all three "say"- "stay" cases, the largest residual performance difference met the crucial criterion for perceptual equivalence ("cooperating cues" > "conflicting cues"). These findings suggest that the "say"- "stay" residual performance was due to perception of the sub-categorical differences as phonetic, rather than as purely auditory (non-phonetic), distinctions. Had the origin of the residual discriminability been purely auditory, the pattern should have followed the "auditory" prediction, or at least should have followed the "temporal" or "spectral" patterns.

Although the residual discrimination performance of "say"- "stay" listeners cannot be explained by between-category phonetic contrasts, it can be explained by within-category distinctions that are nonetheless phonetic (i.e., relevant to allophonic or articulatory variations). The shift in position of the Obtained peaks toward the D (or SW) boundary for all three groups, coupled with the consistent "day" categorization of D (whose F1 onset was like natural "day" and "stay") and the equivocal categorization of d (whose F1 onset was like natural "say"), suggest that residual discrimination was probably based on a distinction such as 'clear /t/ closure' vs. 'inexact (or weak) /t/ closure.' Both gap duration and F1 onset frequency differences apparently provided equivalent information about within-category, as well as between-category, phonetic distinctions.

Conclusions

The two most important points to be made about Experiment 3 are: (1) that the identification and discrimination patterns of the "temporal" and "spectral" listeners differed substantially from the "say"- "stay" results of Experiments 1 and 2; and (2) that the sinewave "say"- "stay" results were essentially identical to the results of the two earlier experiments with synthetic speech. Only the subjects who perceived "say"- "stay" showed a trading relation and perceptual equivalence between the two acoustic cues. Thus, that pattern of identification and 3-way discrimination appears to occur

specifically with perception of speech contrasts. It is not attributable to general psychoacoustic sensitivities or interactions, since it did not appear in the two nonspeech groups, who received the same acoustic information and task demands. The nonspeech listeners focused on only one acoustic dimension (for the most part), and failed to integrate the two into a unitary percept. When the stimuli were perceived as speech, however, gap duration and spectral information were perceptually integrated in a manner that took account of their common origin in speech. Thus, the two cues are integral in speech perception, but separable in auditory perception (cf. Garner & Morton, 1969).

Residual performance patterns on the 3-way oddity test for the "temporal" and "spectral" listeners contrasted sharply with the pattern for the "say"- "stay" listeners in all three experiments. The residual performance patterns of the former two groups were basically consistent with "auditory" predictions, whereas the residual performance pattern of the "say"- "stay" group was consistent with "phonetic" predictions (as were the residual patterns found in Experiments 1 and 2). Discrimination performance on "say"- "stay" contrasts that could not be explained by phonetic category differences seems nonetheless to result from perception of sub-categorical stimulus differences as phonetic, rather than purely auditory.

SUMMARY AND GENERAL CONCLUSIONS

The three experiments present five major findings that bear on the integration of diverse acoustic properties in speech perception. First, there is a trading relation between the two primary acoustic consequences of the articulatory distinction between "say" and "stay." If unequivocal spectral information about the occurrence of a medial /t/ is provided, listeners hear "stay" when the duration of a silent gap between /s/ and the vocalic syllable minimally specifies a stop closure. However, when spectral information provides only equivocal information about an alveolar stop, listeners need stronger evidence for stop closure from another acoustic cue (i.e., longer silent closure gap) in order to perceive "stay."

Second, the two cues for the speech distinction, although from different acoustic dimensions, are perceptually equivalent. They converge upon a unitary phonetic dimension, and provide qualitatively as well as functionally equivalent information about contrastive speech events. Stimulus tokens within a single phonetic category are quite difficult to distinguish perceptually, even though distinct along two different acoustic dimensions ("conflicting cues" comparisons). However, discrimination is comparatively easy when the parameter values for the same two acoustic dimensions are such that the stimuli being discriminated are in different phonetic categories ("cooperating cues").

Third, the qualitative equivalence of the two cues within a single phonetic dimension seems to reflect equivalent perceptual sensitivity to those physical properties of the synthetic speech stimuli. When subjects listen to "say"- "stay" stimuli, the "phonetic" discrimination pattern emerges even under conditions designed to reduce memory demands and eliminate an experimentally-induced "set" to categorize stimuli before discriminating them.

Fourth, trading relations and perceptual equivalence between cues derive specifically from the integrated perception of multiple acoustic properties as phonetic information, and not from psychoacoustic factors. Those perceptual patterns do not occur when listeners perceive the acoustic variations as nonspeech contrasts. Experiment 3 implies that the perceptual integration of diverse acoustic information is determined by what the listener perceives the stimuli to be, much more than it is by raw stimulus characteristics and/or their interactions with basic properties of the auditory system. Several other recent speech perception findings provide converging support for the notion that performance patterns are determined more by the type of stimulus information the subject is listening for than they are by the absolute physical properties of the stimuli. Identification differences occur not only for sinewave speech continua, dependent on whether the stimuli are heard as speech or nonspeech (Bailey et al., 1978; Dorman, 1979), but also differ for speech continua, dependent on the specific phonetic contrast subjects listen for (Carden, Levitt, Jusczyk, & Walley, 1980). Moreover, discrimination performance for speech continua differs substantially, dependent on whether listeners are focusing on phonetic category information or ignoring phonetic categories to focus on purely acoustic properties (Repp, Note 1).

Fifth, the pattern of residual performance on "say"-"stay" discriminations suggests that even within a phonetic category, acoustic variations in the two cues are treated as if they provide phonetic, not simply auditory, information. To our knowledge, this is the first time it has been possible to distinguish empirically between the contributions of auditory and phonetic perception to speech discrimination performance that cannot be explained by phonetic category differences. For this reason, and also because the differences among the comparison types were small, the effect needs replication. The residual discriminability of speech contrasts should most likely reflect phonetic, not auditory, perceptual contributions whenever the acoustic characteristics of the stimuli fall within the range of natural speech variation, and whenever listeners perceive the stimulus properties as information about speech contrasts.

The pattern of perceptual integration of the two cues by the three groups of "say"-"stay" listeners paralleled the pattern of acoustic and articulatory qualities found in natural "say" and "stay" utterances (cf. Experiment 1). That is, "stay" differs from "say" in that only the former word involves a complete linguo-alveolar closure, which results in a longer closure silence and lower F1 onset frequency than found in the latter word. Perceptual integration of the silence and F1-onset cues indicated that listeners had acted as though both cues provided comparable information about whether a complete linguo-alveolar closure had occurred. This pattern of perception-production similarities leads us to agree with the conclusion of Fitch et al. (1980) that trading relations and perceptual equivalence indicate that phonetic perception takes account of the common articulatory origin of diverse cues for a given speech contrast. The perception-production commonalities implied by our results and those of Fitch et al. may suggest that, when listeners attend to the phonetic properties of speech stimuli, they are perceiving articulatory information provided by the acoustic waveform. An excellent discussion of this possibility can be found in Summerfield (1978). Further corroboration for this hypothesis comes from research on the parallel effects of phonetic context on perception and production. A variety of

context effects indicate that phonetic perception takes account of articulatory consequences--e.g., shifting patterns in perception of consonant contrasts parallel the effects of context on the corresponding articulatory gestures (e.g., Mann, 1980; Mann & Repp, 1979a, 1979b; Miller & Liberman, 1979).

The "say"-"stay" results reported in this paper appear robust, and reflect the perceptual integrity of the multiple acoustic consequences of articulatory gestures as phonetic information. But the possibility that the trading relation/perceptual equivalence pattern may be unique to speech needs further investigation. To learn whether that perceptual pattern is uniquely human, the responses of animals to speech and nonspeech contrasts conveyed by multiple physical cues (even nonauditory) might be studied (cf. Kuhl, 1978; Liberman & Pisoni, 1977; but also compare Kuhl & Miller, 1975, 1978; Morse & Snowden, 1975; Waters & Wilson, 1976). For example, a recent report of discrimination among leaf categories by pigeons (Cerella, 1979) suggested that the animals may have treated several dimensions of contrast among leaf outlines as equivalent (e.g., smooth vs. serrated edge, shallow vs. deep notches between lobes, etc.) in their perception of natural classes of leaves (oak vs. non-oak). Further research would be necessary, however, to determine the completeness of the pigeon's perceptual integration of leaf-outline dimensions--that is, to determine whether they would show trading relations and perceptual equivalence among the diverse features.

Also, to assess whether perceptual equivalence is uniquely characteristic of speech perception, or whether it may be a more general quality in perception of complex acoustic information for naturally-occurring contrastive events, the perceptual integration of multiple cues might be explored for familiar nonspeech events that are rich in dynamic acoustic information. For example, there are probably spectral as well as temporal contrasts between the acoustic products of plucking vs. bowing actions on a violin string (Schelleng, 1973), or between the acoustic consequences of hard vs. soft attack in the playing of piano notes (Weyer, 1976, 1976/1977). Contrastive nonspeech properties such as these might also be perceptually integrated, but it is not clear a priori whether such integration would imply functional and qualitative equivalence among the diverse acoustic cues. Answers to questions about multiple acoustic properties of natural nonspeech events, and about perceptual integration of those (possible) properties, still await empirical exploration.

The strength of the current findings implies that perceptual equivalence of multiple cues to a given phonemic contrast is a key aspect of adult speech perception. Developmental research on trading relations and perceptual equivalence in speech perception may aid in understanding the interplay of maturation, perceptual experience, and articulatory competence in the ontogeny of the general ability to perceive phonetically-relevant characteristics of human speech (again, see discussions by Kuhl, 1978; Liberman & Pisoni, 1977). Such research would help in appraising whether certain acoustic contrasts elicit innate or biologically-determined perceptual responses, while others gain an effect on perception primarily through receptive and productive language experience. For example, the voiced-voiceless distinction can be cued for adults by contrasts in either VOT or F1 onset frequency (e.g., Lisker, 1975). However, though young children and even very young infants respond to VOT categories much like adults (cf. Jusczyk, 1980; Kuhl, 1978), children do not respond strongly to F1 onset distinctions until they are

around five years old (Simon & Fourcin, 1978), at which time they may show phonetic trading relations that are smaller in magnitude than the corresponding adult trading relations (Robson, Morrongiello, Best, & Clifton, Note 2). These facts may suggest that the development of trading relations in phonetic perception is dependent on fairly extensive language experience, and would begin to show up only during the preschool-kindergarten years. On the other hand, a recent study of 6-month-olds' perception of the /slIt/-/splIt/ contrast, cued either by the artificial introduction of silence alone or by natural silence plus /p/ bursts and transitions, suggests that even young infants might show some evidence of a phonetic trading relation (Morse, Eilers, & Gavin, Note 3).

REFERENCE NOTES

1. Repp, B. H. Influence of vocalic context on perception of the [ʃ]-[s] distinction: IV. Fricative noise discrimination. Manuscript in preparation, Haskins Laboratories, 1980.
2. Robson, R., Morrongiello, B., Best, C. T., & Clifton, R. K. Trading relations in the perception of speech by five-year-olds and adults. Paper presented at meeting of the Eastern Psychological Association. Hartford, Connecticut, April, 1980.
3. Morse, P. A., Eilers, R., & Gavin, W. Exploring the perception of the "sound of silence" in early infancy. Paper presented at the 2nd meeting of the International Conference on Infant Studies. New Haven, Connecticut, April, 1980.

REFERENCES

- Abramson, A. S., & Lisker, L. Voice onset time in stop consonants: Acoustic analysis and synthesis. In D. E. Commins (Ed.), Proceedings of the 5th Congress of International Acoustics. Liege, France: Imp. G. Thone, 1965.
- Bailey, P. J., & Summerfield, A. Q. Some observations on the perception of s + stop clusters. Haskins Laboratories Status Report on Speech Research, 1978, SR-53, 25-60.
- Bailey, P. J., Summerfield, Q., & Dorman, M. F. On the identification of sinewave analogues of certain nonspeech sounds. Haskins Laboratories Status Report on Speech Research, 1977, SR-51/52, 1-25.
- Carden, G., Levitt, A. G., Jusczyk, P. W., & Walley, A. Evidence for phonetic processing of cues to place of articulation: Perceived manner affects perceived place. Haskins Laboratories Status Report on Speech Research, 1980, SR-62, this issue.
- Cerella, J. Visual classes and natural categories in the pigeon. Journal of Experimental Psychology: Human Perception and Performance, 1979, 5, 68-77.

- Creelman, C. D., & MacMillan, N. A. Auditory phase and frequency discrimination: A comparison of nine methods. Journal of Experimental Psychology: Human Perception and Performance, 1979, 5, 146-156.
- Crowder, R. G. The sound of vowels and consonants in immediate memory. Journal of Verbal Learning and Verbal Behavior, 1971, 10, 587-596.
- Crowder, R. G. Representation of speech sounds in precategorical acoustic storage. Journal of Experimental Psychology, 1973, 1, 14-24.
- Crowder, R. G., & Morton, J. Pre-categorical acoustic storage (PAS). Perception & Psychophysics, 1969, 5, 365-373.
- Cutting, J. E., & Rosner, B. S. Categories and boundaries in speech and music. Perception & Psychophysics, 1974, 16, 564-570.
- Cutting, J. E., & Rosner, B. S. Discrimination functions predicted from categories in speech and music. Perception & Psychophysics, 1976, 20, 87-88.
- Cutting, J. E., Rosner, B. S., & Foard, C. F. Perceptual categories for musiclike sounds: Implications for theories of speech perception. Quarterly Journal of Experimental Psychology, 1976, 28, 361-378.
- Darwin, C. J., & Baddeley, A. D. Acoustic memory and the perception of speech. Cognitive Psychology, 1974, 6, 41-60.
- Delattre, P. C., Liberman, A. M., & Cooper, F. S. Acoustic loci and transitional cues for consonants. Journal of the Acoustical Society of America, 1955, 27, 679-773.
- Divenyi, P. L. Some psychoacoustic factors in phonetic analysis. Proceedings of the 9th International Congress of Phonetic Sciences. Copenhagen: Stougaard Jensen, 1979.
- Divenyi, P. L., & Danner, W. F. Discrimination of time intervals marked by brief acoustic pulses of various intensities and spectra. Perception & Psychophysics, 1977, 21, 125-142.
- Divenyi, P. L., & Sachs, R. M. Discrimination of time intervals bounded by tone bursts. Perception & Psychophysics, 1978, 24, 429-436.
- Dorman, M. F. On the identification of sinewave analogues of CV syllables. Proceedings of the Ninth International Congress of Phonetic Sciences, Vol. II. Copenhagen: University of Copenhagen, 1979, 453-461.
- Dorman, M. F., Raphael, L. J., & Isenberg, D. Acoustic cues for fricative/affricate contrast in word-final position. Haskins Laboratories Status Report on Speech Research, 1979, SR-58, 217-230.
- Dorman, M. F., Raphael, L. J., & Liberman, A. M. Some experiments on the sound of silence in phonetic perception. Journal of the Acoustical Society of America, 1979, 65, 1518-1532.
- Dorman, M. F., Studdert-Kennedy, M., & Raphael, L. J. Stop consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. Perception & Psychophysics, 1977, 22, 109-122.
- Fant, C. G. M. Descriptive analysis of the acoustic aspects of speech. Logos, 1962, 5, 3-17.
- Fitch, H. L., Halwes, T. G., Erickson, D. M., & Liberman, A. M. Perceptual equivalence of two acoustic cues for stop consonant manner. Perception & Psychophysics, 1980, 27, 343-350.
- Fujisaki, H., & Kawashima, T. On the modes and mechanisms of speech perception. Annual Report of the Audio Engineering Society, 1969, 28, 67-73.
- Garner, W. R., & Morton, J. Perceptual independence: Definitions, models, and experimental paradigms. Psychological Bulletin, 1969, 72, 233-259.
- Green, D. M. An introduction to hearing. New York: John Wiley & Sons, 1976.

- Haggard, M. P., Ambler, S., & Callow, M. Pitch as a voicing cue. Journal of the Acoustical Society of America, 1970, 47, 613-617.
- Harris, K. S., Hoffman, H. S., Liberman, A. M., Delattre, P. C., & Cooper, F. S. Effect of third-formant transitions on the perception of the voiced stop consonants. Journal of the Acoustical Society of America, 1958, 30, 122-126.
- Hoffman, H. S. Study of some cues in the perception of the voiced stop consonants. Journal of the Acoustical Society of America, 1958, 30, 1035-1041.
- Jusczyk, P. W. Infant speech perception: A critical appraisal. In P. D. Eimas & J. A. Miller (Eds.), Perspectives on the study of speech. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1980, in press.
- Kaplan, H. L., MacMillan, N. A., & Creelman, C. D. Methods & designs: Tables of d' for variable-standard discrimination paradigms. Behavior Research Methods & Instrumentation, 1978, 10, 796-813.
- Kuhl, P. K. Predispositions for perception of speech-sound categories: A species-specific phenomenon? In F. D. Minifie & L. L. Lloyd (Eds.), Communicative and cognitive abilities -- Early behavioral assessment. Baltimore, Md.: University Park Press, 1978.
- Kuhl, P. K., & Miller, J. D. Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. Science, 1975, 190, 69-72.
- Kuhl, P. K., & Miller, J. D. Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. Journal of the Acoustical Society of America, 1978, 63, 905-917.
- Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. Perception of the speech code. Psychological Review, 1967, 74, 430-460.
- Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. The discrimination of speech sounds within and across phoneme boundaries. Journal of Experimental Psychology, 1961, 61, 379-388.
- Liberman, A. M., & Pisoni, D. B. Evidence for a special speech-perceiving subsystem in the human. In T. H. Bullock (Ed.), Recognition of complex acoustic signals. Berlin: Dahlem Konferenzen, 1977.
- Liberman, A. M., & Studdert-Kennedy, M. Phonetic perception. Haskins Laboratories Status Report on Speech Research, 1977, SR-50, 21-60.
- Lisker, L. Is it VOT or a first-formant transition detector? Journal of the Acoustical Society of America, 1975, 57, 1547-1551.
- Lisker, L. Rapid vs. rabid: A catalogue of acoustic features that may cue the distinction. Haskins Laboratories Status Report on Speech Research, 1978, SR-54, 127-132.
- Lisker, L., Liberman, A. M., Erickson, D. M., Dechovitz, D., & Mandler, R. On pushing the voice-onset-time (VOT) boundary about. Language and Speech, 1977, 20, 209-216.
- MacKain, K. S., Best, C. T., & Strange, W. Native language effects on liquid perception. Journal of the Acoustical Society of America, 1980, 67 (Suppl.), S27 (Abstract, K12).
- MacMillan, N. A., Kaplan, H. L., & Creelman, C. D. The psychophysics of categorical perception. Psychological Review, 1977, 84, 452-471.
- Mann, V. A. Influence of preceding liquids on stop-consonant perception. Journal of the Acoustical Society of America, 1980, 67 (Suppl.), S99 (Abstract, QQ1).

- Mann, V. A., & Repp, B. H. Influence of vocalic context on perception of the [f]-[s] distinction: I. Temporal factors. Haskins Laboratories Status Report on Speech Research, 1979, SR-57, 49-64. (a)
- Mann, V. A., & Repp, B. H. Influence of preceding fricative on stop consonant perception. Haskins Laboratories Status Report on Speech Research, 1979, SR-57, 65-82. (b)
- McGovern, K., & Strange, W. The perception of /r/ and /l/ in syllable-initial and syllable-final position. Perception & Psychophysics, 1977, 21, 162-170.
- Miller, J. A., & Liberman, A. M. Some effects of later-occurring information on the perception of stop consonant and semivowel. Perception & Psychophysics, 1979, 25, 457-465.
- Miller, J. D., Wier, C. C., Pastore, R. E., Kelly, W. J., & Dooling, R. J. Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception. Journal of the Acoustical Society of America, 1976, 60, 410-417.
- Morse, P. A., & Snowden, C. T. An investigation of categorical speech discrimination by rhesus monkeys. Perception & Psychophysics, 1975, 17, 9-16.
- Pastore, R. E., Ahroon, W. A., Baffuto, K. J., Friedman, C., Puleo, J. S., & Fink, E. A. Common factor model of categorical perception. Journal of Experimental Psychology: Human Perception and Performance, 1977, 4, 686-696.
- Pastore, R. E., Friedman, C. J., & Baffuto, K. J. A comparative evaluation of the AX and two ABX procedures. Journal of the Acoustical Society of America, 1976, 60 (Suppl.), S120. (Abstract)
- Pisoni, D. B. On the nature of categorical perception of speech sounds. Haskins Laboratories Status Report on Speech Research, 1971, SR-27, 209-210.
- Pisoni, D. B. Auditory short term memory and vowel perception. Memory & Cognition, 1975, 3, 7-18.
- Pisoni, D. B. Identification and discrimination of the relative onset time of two-component tones: Implications for voicing perception in stops. Journal of the Acoustical Society of America, 1977, 61, 1352-1361.
- Pisoni, D. B., & Lazarus, J. H. Categorical and noncategorical modes of speech perception along the voicing continuum. Journal of the Acoustical Society of America, 1974, 55, 328-333.
- Pisoni, D. B., & Tash, J. Reaction times to comparisons within and across phonetic boundaries. Perception & Psychophysics, 1974, 15, 285-290.
- Repp, B. H. Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants. Language and Speech, 1979, 22, 173-189.
- Repp, B. H., Healy, A. F., & Crowder, R. G. Categories and context in the perception of isolated steady-state vowels. Journal of the Acoustical Society of America, 1979, 5, 129-145.
- Repp, B. H., Liberman, A. M., Eccardt, T., & Pesetsky, D. Perceptual integration of temporal cues for stop, fricative, and affricate manner. Journal of Experimental Psychology: Human Perception and Performance, 1978, 4, 621-637.
- Repp, B. H., & Mann, V. A. Perceptual assessment of fricative-stop coarticulation. Journal of the Acoustical Society of America, 1980, 67 (Suppl.), S100 (Abstract, QQ8).

- Schelleng, J. C. The bowed string and the player. Journal of the Acoustical Society of America, 1973, 53, 26-41.
- Simon, C., & Fourcin, A. J. Cross-language study of speech-pattern learning. Journal of the Acoustical Society of America, 1978, 63, 925-935.
- Stevens, K. N. The role of rapid spectrum changes in the production and perception of speech. In L. L. Hammerich & R. Jakobson (Eds.), Form & substance: Festschrift for Eli Fischer-Jørgensen. Copenhagen: Akademisk Forlag, 1971.
- Stevens, K. N. The quantal nature of speech: Evidence from articulatory-acoustic data. In L. Pinson & D. Denes (Eds.), Human communication: A unified view. Cambridge, Mass.: M.I.T. Press, 1974.
- Studdert-Kennedy, M. Speech perception. In N. J. Lass (Ed.), Contemporary issues in experimental phonetics. New York: Academic Press, 1976.
- Summerfield, A. Q. Perceptual learning and phonetic perception. Interrelations of the communicative senses. Proceedings of the NSF conference at Asilomar. Washington, D.C.: NSF Publication, 1978.
- Summerfield, A. Q., & Haggard, M. P. On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. Journal of the Acoustical Society of America, 1977, 62, 435-448.
- Waters, R. S., & Wilson, W. A. Speech perception by rhesus monkeys: The voicing distinction in synthesized labial and velar stop consonants. Perception & Psychophysics, 1976, 19, 285-289.
- Weyer, F. D. Time-frequency-structures in the attack transients of piano and harpsichord sounds - I. Acustica, 1976, 35, 233-252.
- Weyer, R. D. Time-varying amplitude-frequency-structures in the attack transients of piano and harpsichord sounds - II. Acustica, 1976-77, 36, 241-252.

FOOTNOTES

¹The use of these words might have introduced a lexical bias (frequency counts for "say" and "stay" are not equal), although the effect of that bias on performance could not have interfered with our test results, since lexical bias should have involved both "say"- "stay" continua equally. At any event, lexical bias is unlikely, given that our early stimulus development was based on the /se/-/ste/ contrast. Neither pilot stimulus is a word, and neither is phonologically permissible in American English because of the final lax vowel; nonetheless, a trading relation was found between silence and F1 onset as cues for the /se/-/ste/ contrast.

²The /t/ in "stay" is unaspirated ([t] as opposed to [t^h]). When an unaspirated /t/ follows an /s/ in American English, as in "stay," it is identified in that context as "t" ("stay"). However, if the /s/ is removed from "stay" (leaving a vocalic syllable similar to our synthetic stimuli with low F1 onset frequency), the isolated [teɪ] is identified as "day" rather than "tay" because word-initial voiceless stops in spoken American English are nearly always aspirated (i.e., [t^h]).

³Pilot testing had indicated that subjects would need an additional 24 msec of silent gap (approximately) to begin hearing "stay" for the d continuum, relative to the D continuum.

⁴Although 12-18 judgments per comparison are typically obtained in discrimination tests, we used only six judgments per comparison to keep the total testing time short. We wanted to compare identification and discrimination results gathered within a single test session, to minimize possible changes in response criteria, attention level, etc., that might occur between days or within long test sessions. In addition, since we were most interested in the subjects' "natural" or "normal" perception of the stimuli, we kept presentations per stimulus at the smallest number likely to yield reliable response functions. These concerns also related to our use of 10, rather than the usual 20, judgments per token in the identification test. Because the number of judgments per item in both tests was low, the potential for response variability was higher than usual. This situation thus provides a conservative test of our hypotheses.

⁵The Stimulus Pairs simple effect for the "conflicting cues" comparisons did not support a performance peak near the category boundary; instead, it indicated lower discrimination performance for the extreme endpoints from the two continua [s(0)D vs. s(24)d, s(8)D vs. s(32)d, and s(112)D vs. s(136)d] than for all other comparisons.

⁶Although the oddity task is commonly assumed to have higher memory demands than other discrimination procedures used in auditory research, and hence to yield the lowest performance level, no direct comparisons of oddity versus other tasks are known (personal communications: Pastore; Remez; Repp), with one recent exception. MacKain, Best, & Strange (1980) found slightly higher above-chance performance for an AXB than an oddity task, using a synthetic /rak/-/lak/ continuum; this corroborates common intuitions about the relative difficulty (memory demands) of the oddity task.

⁷Throughout this methodological discussion, we are referring to variable-standard rather than fixed-standard discrimination designs. Although fixed-standard designs yield higher performance than variable-standard designs, according both to theoretical models (SDT analysis: MacMillan et al., 1977) and to empirical work with nonspeech stimuli (Creelman & MacMillan, 1979), those task differences may be very small for discrimination of speech (Repp, Note 1). Furthermore, we were constrained to use a variable-standard design in order to look at discriminability of stimulus pairs throughout the range of the two stimulus continua, and the 3-way discrimination design dictated that the magnitude of within-pair differences be fixed.

⁸True d' values from 2IAX data are equal to or slightly higher than ABX or 4IAX d' values for frequency discriminations, and may be even more improved for phase discriminations, relative to the latter two paradigms. In Creelman & MacMillan (1979, pp. 151-152), compare variable-standard designs in Figure 4 with the interrelation of Figure 1 and Table 2, which show the same pattern of ABX-4IAX performance level differences found in speech discrimination tests (e.g., Pisoni, 1971, 1975).

⁹Thanks to Philip Rubin for his helpful modifications of the sinewave synthesis program originally written by Rod McGuire at Haskins Laboratories.

¹⁰The two sinewave syllables lacked the final diphthongization from /e/ to /i/ because the additional frequency modulation seemed to make their

speechlike qualities too obvious, and were physically more analogous to /de/ than to /dei/. However, the offset of the three sinewaves was such that they sounded like "day" rather than "deh" to most listeners who heard them as speech.

¹¹It was especially important in Experiment 3 to keep testing time and stimulus repetition low, to minimize the possibility that nonspeech listeners might spontaneously begin to hear the stimuli as speech after prolonged exposure, and shift their perceptual behavior.

¹²As these descriptions of the subgroups suggest, there was great variation in individual perceptions of these stimuli, and attempts to impose perceptual characterizations did not work consistently. For example, some subjects were told to listen for "say" and "stay," but failed to hear the words; other subjects were told that the stimuli were "hiss + gap + chime," yet several heard words, including "say" and "stay." Both the individual variation and the inconsistent response to "perceptual instruction" seem to be characteristic of tests with sinewave speech analogues, having been noted before (e.g., Bailey et al., 1978; Summerfield, personal communication).

¹³The functions appeared a bit rough because of the small number of subjects, the difficulty that three of the subjects had in keeping 'tuned' to "say" and "stay" throughout the test, and especially because of the subject who reported listening for different water drips through most of the 3-way oddity test; all of these factors make interpretation of any visual irregularities difficult. For example, there is an apparent bimodality in the "cooperating cues" peak, but the 'dip' between the two highest points represents a total drop of only 4-5 correct responses. Furthermore, the "double-peak" pattern was shown by only two subjects, each of whom showed a 'dip' of 2 responses in magnitude. One of these subjects was a 'spontaneous' "say"- "stay" listener, and had the noisiest data of this group.

¹⁴Although the Obtained and Predicted data showed the same order of "peak-range" performance levels near the category boundary, the Obtained boundary-related peak (20 msec) was shifted from the Predicted peak (36 msec), and there was a second unexpected Obtained peak (60-68 msec). Both deviations suggest that the oddity and AXB tests may have tapped different perceptual processes for this group; they may have been using three categories, rather than two. The shallow slopes of the AXB functions suggest that the two prototype categories we presented (A and B) might not have been the most appropriate for these listeners. The 20 msec peak hints that one perceived contrast may have been 'contiguity between hiss and sinewave' vs. 'delay between hiss and sinewave,' which would be consistent with the general psychoacoustic boundary at 20 msec for detection of temporal differences between components of two-part signals (e.g., Miller et al., 1977; Pastore et al., 1977; Pisoni, 1977). The later peak (60-68 msec) may distinguish 'a 2-component signal' vs. 'two separate signals.' A one-way ANOVA on the three Comparison Types, for mean "peak-range" performances around the second peak (56-88 msec), showed nonsignificant differences among the three comparison types, indicating that in that range the listeners must have responded to SW-SW differences in all three comparison types. However, the drop in performance between 72-96 msec makes it unlikely that the "temporal" listeners had merely focused on the spectral contrasts at longer gap durations. It is more

likely that they were still attending to temporal information, and that sensitivity to gap changes in that range may have been differentially affected by the SW vs. sw onset spectra (and/or discrimination as a function of gap duration is nonmonotonic).

Appendix A

In both stimuli, there was a 25 msec risetime, and 50 msec falltime; the parameter values for the amplitudes of F2 and F3 throughout the stimuli were 4/5 that of F1. The F0 contour began at 120 Hz and remained at that frequency for 240 msec, after which it fell linearly to 90 Hz during the final 50 msec. F3 began with a linear 40 msec transition from 3196 Hz to 2694 Hz (somewhat exaggerated with respect to natural stimuli), and remained at the latter frequency for 130 msec, after which it rose linearly to 3029 Hz during the ensuing 70 msec. F2 remained at 1840 Hz for the initial 150 msec, and rose linearly to 2298 Hz during the following 90 msec. F2 and F3 remained at their last-named frequencies during the final 50 msec. F1 reached 611 Hz at the end of the initial 40 msec transitions, remained at that frequency for 110 msec, then fell linearly to 304 Hz during the final 90 msec.

Appendix B

We used the following formula to predict each subject's performance for each of the 44 stimulus comparisons tested (from McGovern & Strange, 1977):

$$P_{\text{corr}} = (1 + 2[P_a - P_b]^2)/3,$$

where P_{corr} is Predicted probability of correct responses for a given stimulus comparison, P_a is Obtained proportion of "stay" responses to stimulus a, and P_b is the observed proportion of "stay" responses to stimulus b in the comparison. Chance level responding was 33%.

Appendix C

To determine true d' values for Obtained and Predicted data in the 2IAX test, it was necessary to calculate $P[H]$ (probability of a hit) and $P[FA]$ (probability of a false alarm). These were derived by the following formulae:

$$P[H] = P("S"/S)$$

for the Obtained hit rate, where P is 'probability' or 'proportion,' "S" is '"same" responses,' and S is 'given a physically-identical AX pair';

$$P[H] = P("stay"/A)^2 + P("stay"/X)^2$$

for the Predicted hit rate, where $P("stay"/A)$ is 'proportion of "stay" responses to stimulus A in the identification test,' and likewise for stimulus X in $P("stay"/X)$;

$$P[FA] = P(D/S)$$

for the Obtained false alarm rate, where "D" is '"different" responses'; and,

$$P[FA] = P("stay"/A) \cdot P("say"/A) + P("stay"/X) \cdot P("say"/X)$$

for Predicted false alarm rate. These values of $P[H]$ and $P[FA]$ were used to look up true d' values in the table provided by Kaplan et al. (1978). Because this table (and the computational formula used to derive the tabled values) does not allow for $P[H]$ and $P[FA]$ of either 0.00 or 1.00, we substituted 0.01 and 0.99, respectively, for occurrences of those values in our data. Thus, the true d' values were artificially constrained (although to a small extent) at the high and low extremes ($d'_{\text{max}} = 6.93$; $d'_{\text{min}} = 0.00$).

DUPLEX PERCEPTION OF ACOUSTIC PATTERNS AS SPEECH AND NONSPEECH

Alvin M. Liberman⁺ and David S. Isenberg⁺⁺

Abstract. If certain acoustic elements of a synthetic syllable are isolated and presented separately to the two ears, listeners fuse them into a coherent phonetic percept, while also perceiving one element as a nonspeech 'chirp'. This duplex perception is used to provide evidence for a mode of perception that is distinctively phonetic.

A biological specialization for language could become manifest in several ways. At the level of speech perception, it might appear as a distinctive mode (or subsystem), different from the auditory mode(s), and comprising just those processes that come into play when, on being presented with the sounds of speech, a listener perceives consonants and vowels. That such a mode--let us call it 'phonetic'--may, in fact, exist is implied by a variety of circumstantial evidence (1). Our aim here is to offer evidence of a more direct sort. To that end, we take advantage of a procedure, contrived by Rand (1974) (2), that produces an apparently odd, and possibly unique, phenomenon: two quite different percepts arise simultaneously when the acoustic constituents of a (synthetic) syllable are separated and presented dichotically. One of these percepts is a coherent syllable that sounds just like the original, having been formed, evidently, by a fusion of the dichotically presented constituents. The other percept is a nonspeech 'chirp' that contrasts strikingly with the syllable and corresponds closely to what one of the constituents sounds like in isolation. We take this duplex percept to be of interest, not only because it demonstrates the simultaneous operation of phonetic and auditory modes, but also because it can be used to determine why, in the perception of speech, certain aspects of the signal are effective cues.

In the experiment reported here we tested for duplex perception in the case of phonetic structures that contrast the presence of one or another stop consonant ([spal or [ɛtal]) with its absence ([sal]), and, at the same time, attempted to find out why, for the perception of those contrasts, an appropriately positioned interval of silence should be important. In the top

⁺Also at the University of Connecticut and Yale University.

⁺⁺Currently a postdoctoral fellow at the Massachusetts Institute of Technology.

Acknowledgment. This work was supported by grant HD 01994 from the National Institute of Child Health and Human Development and by grants NS-05493 and NS-07040 from the National Institutes of Health. We thank Terry Halwes for help with the stimuli. For useful comments we thank Michael Studdert-Kennedy, Bruno Repp, and Dennis Klatt.

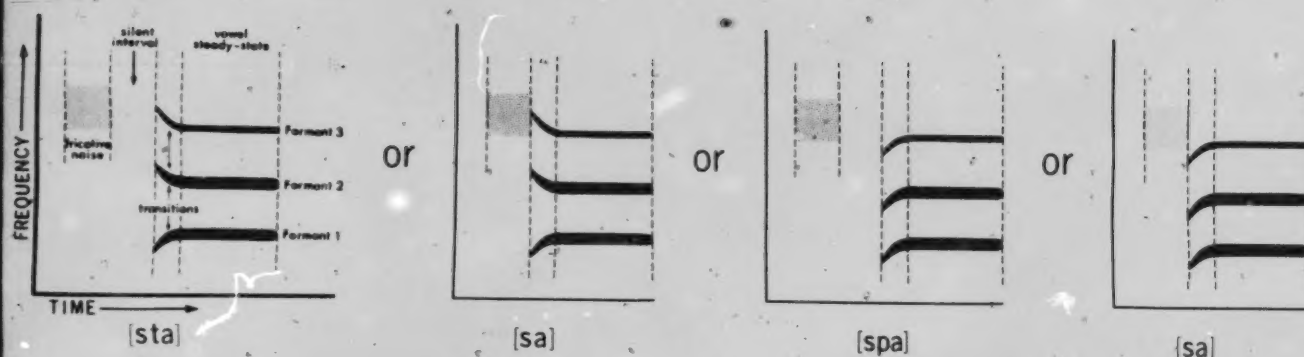
half of Figure 1 we see the acoustic features that are of interest. For the distinction between [spa] and [sta], transitions of the second and third formants can be sufficient cues, but either of these syllables can be reduced to [sa] by removing the silence between the noise and the vocalic section. Thus, given appropriate formant transitions, the presence or absence of a short period of silence can determine whether or not those transitions specify the stop consonants [p] and [t] (3).

To understand the importance of silence as a cue for stop consonants, we can invoke either of two quite different accounts: one auditory, the other phonetic. An auditory account might have it, for example, that the silence enables the transition cues to evade (forward) auditory masking by the fricative noise. Failing that, it could fall back on the assumption that the silence and the transitions engage in an auditory interaction of some unspecified sort, from which emerges the percept we know as a stop consonant. But the particular nature of an auditory account need not concern us. For our purposes, the important characteristic of such a view is that the effect of silence would occur for all sounds that have certain attributes, and it would not matter that they were, or were not, perceived as speech.

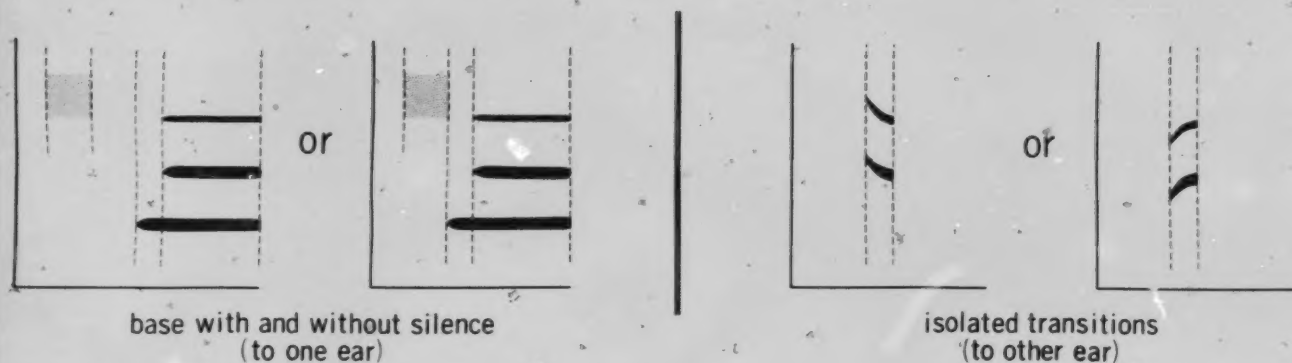
According to a phonetic interpretation, on the other hand, the stop-producing effect of silence reflects the operation of a distinctive process, specialized to perceive the silence in a linguistically appropriate way. Such a process would treat the presence or absence of silence as phonetically relevant information, revealing either that the talker's vocal tract closed, as it must to produce the stop consonant in [spa] or [sta], or that it did not, as it does not when the talker articulates the "stopless" [sa]. On this interpretation, silence would have its effect only when the sounds were being processed as speech.

The phonetic interpretation of the role of silence has found support in several recent experiments (Repp, Liberman, Eccardt, & Pesetsky, 1978; Bailey & Summerfield, in press; Fitch, Halwes, Erickson, & Liberman, 1980), including one (Dorman, Raphael, & Liberman, 1979) that is most directly relevant to our immediate concern. There, the investigators found that a preceding silent interval was, indeed, necessary if formant transitions were to function as cues for stop consonants, but that silence had little or no effect on the identification of those same transitions when, having been isolated from the rest of the pattern, they were perceived as nonspeech chirps. This result is, of course, open to the reservation that the transition cues were, after all, in different acoustic contexts in the two conditions of the experiment. The present experiment was designed to speak to that reservation. To that end, we exploit the phenomenon of duplex perception to determine whether the silence affects the speech percept, but not the corresponding nonspeech, when the acoustic patterns that produce both percepts are exactly the same.

To engineer the duplex percept, we separated the synthetic speech patterns of the top half of Figure 1 into two parts, as shown in the bottom half, and presented them dichotically, one part to one ear and one part to the other. One of the parts, shown at the right of the figure, is simply the transitions of the second and third formants (4), the only cues in these patterns for the perceived difference between [spa] and [sta]. When presented in isolation, as they might be to one ear alone or to both ears equally, these



NORMAL (BINAURAL) PRESENTATION



DUPLEX-PRODUCING (DICHOTIC) PRESENTATION

Figure 1. Schematic representation of the stimuli.

transitions are perceived variously. Some listeners identify them as speechlike sounds of some kind (5), but not as [p] or [t]; others hear them as musical glissandos; most listeners--all in our experiment--find 'chirps' an acceptable characterization. These 'chirps' sound less different in isolation than when they fuse with the fixed syllabic context to cue a phonetic contrast (Mattingly, Liberman, Syrdal, & Halwes, 1971); nevertheless, they are discriminable.

The other part of each to-be-dichotically-presented pair is the pattern labeled 'base' at the left of the figure (6). This is what remains of the normal patterns when the second- and third-formant transition cues have been removed and the transition of the first formant straightened (7). By itself, the base is perceived as a syllable--that is, as speech--but a free identification test revealed that it was usually heard as [sa] or [s?a] or [s-a], sometimes as [sta], never as [spa] (8).

These patterns were used experimentally to answer two questions. (1) Do listeners fuse the isolated transitions with the base to perceive an appropriate syllable ([spa]), [sta]), while simultaneously perceiving the isolated transitions as nonspeech chirps? (2) Does the silence cue affect the perception of the transitions only when they are being perceived as speech?

The procedure was as follows. A single 'trial' consisted always of one dichotically presented pair--base to one ear, isolated transitions to the other--followed after 420 msec by another such pair (9). The isolated transitions were presented at a time corresponding to their occurrence in the normal (binaural) case. On each trial, the base component of one of the dichotic pairs had the silent interval appropriate for a stop consonant, the other did not; both orders of presentation of the two conditions of the base (silence, no silence) were used. The pairing of the base component and the isolated transitions (appropriate for [p] or for [t]) exhausted all possible combinations, yielding 8 types of trials. Thus, a listener could not know which of the two base components would be in either dichotic pair, nor could he know which of the two transition cues was to be paired with it. For presentation to the subjects in the experiment, 10 tokens of each of the 8 types of trial were recorded, in random fashion, onto a test tape (10).

The subjects, ten college students who were naive with respect to the nature of the experiment, listened three times to the test tape (11). First, they were asked to attend to the speech side of the percept, and, on each trial, to identify the pair as [spa]-[sa], [sta]-[sa], [sa]-[spa], or [sa]-[sta]. This was to see if they correctly fused the isolated transitions with the base, and also if the silence cue affected the transitions when, having been fused with the base, those transitions were perceived as speech. Next, the subjects were instructed to attend to the nonspeech side of the percept, and, on each trial, to judge the chirps as 'same' or 'different'. This was to make sure that the subjects did also perceive the transitions as nonspeech chirps, and to see whether or not that perception was affected by the silence cue. Finally, subjects were asked to attend, on each trial, to both sides of the duplex percept, and to respond as before (12). This was to test the possibility that the responses given previously might have depended on the subjects' being able, by an effort of attention, to experience but one side of the percept at a time.

The results are presented in the table. Looking first at the judgments made about the speech side of the duplex percept, we see that all subjects correctly identified the syllables at better than a 1% level of significance. This was so, not only when the subjects were attending only to the speech side, but also when they had to attend to (and report on) both sides. Plainly, the two kinds of isolated transitions did fuse with the base to produce the appropriate stop consonants. Just as plainly, the silence cue grossly affected the perception of the transitions when, on the speech side of the duplex percept, those transitions were being used as cues for stop consonants: when silence was paired with the transitions, subjects perceived (appropriately) the stops in [spa] and [sta]; in the absence of the silence cue, they perceived the 'stopless' [sa].

On the nonspeech side, we should comment first that all the subjects did hear the transitions as chirps; moreover, they heard them whether or not there was silence in the base. Looking at the table, we see that most of the subjects judged accurately that the chirps were the same or different according to whether the transitions that gave rise to them were the same or different. When they had to judge but one of the two sides of the duplex percept, seven of the ten subjects performed correctly at better than a 1% level of significance (13); when both judgments were required on each trial, then eight of the ten met that criterion (14). Given that most of the subjects performed 'correctly' in regard to the chirps, we note especially that they perceived the chirps to be the same when the transitions that produced them were the same, though one of the transitions had been paired with silence and the other not (15). We see, thus, on the nonspeech side of the duplex percept that silence did not affect perception of the transitions.

In comparing the results for the two sides of the duplexity, we should emphasize that perception of the syllable on the speech side is just like perception of that syllable when, as is normal, base and transitions are combined and presented to both ears. In neither case does a listener consciously perceive a chirp and then consciously decide whether or not to interpret it as a stop consonant. Rather, he hears simply [spa], [sta], or [sa]. Thus, given a trial in which the two stimulus patterns--the one with silence and the one without--had the same t-type transitions, for example, the subjects could not have judged, as they did with the chirps, that the percepts directly corresponding to the transitions were the same or different. What they could judge, and did, was that one of the patterns was [sta], the other [sa]. In effect, then, changing the silence cue caused identical transitions to be perceived as different. But only on the speech side. On the nonspeech side, and under identical circumstances, those transitions were perceived to be the same.

We draw two conclusions from these results. First, duplex perception did occur. Given, in one ear, a base component that was perceived in isolation as a syllable that did not contain [p] or [t], and given, in the other ear, isolated transitions that were perceived in isolation as not-very-speechlike chirps, there were three consequences in perception: (1) the two components (base and transitions) fused to produce the same syllable--[spa], [sta], or [sa]--that would have been produced had the components been electronically mixed and presented in normal fashion; (2) the transitions that, in isolation, produced nonspeech chirps continued to do so; and (3) the base component that,

Table 1

Performance on the speech-identification and chirp-discrimination tasks. Chance performance was 25% (one of four pairs of responses) on the former, 50% (one of two responses) on the latter. Asterisk indicates above-chance performance, according to a binomial test, at the level of $p < .01$.

Percent Correct

Subject	Each task separately		Both tasks together	
	Speech	Chirp	Speech	Chirp
1	88.8*	98.8*	71.3*	90.0*
2	100.0*	78.8*	98.8*	70.0*
3	98.8*	92.5*	82.5*	71.3*
4	88.8*	65.0	73.8*	80.0*
5	100.0*	88.8*	100.0*	70.0*
6	93.8*	72.5*	82.5*	71.3*
7	97.5*	50.0	100.0*	57.5
8	93.8*	50.0	77.5*	62.5
9	90.0*	90.0*	92.5*	86.3*
10	91.3*	93.8*	56.3*	68.8*

in isolation, produced a syllable different from the "correct" fused syllable was not perceived as such, having been absorbed, as it were, in the process of fusion (14). What the listeners heard, then, was precisely duplex: a syllable in which the transition cues were interpreted in phonetically normal fashion, and a nonspeech chirp that resembled the percept produced when the transitions are presented alone. Evidently, the brain can simultaneously process the same acoustic pattern in two different ways. We have termed the one way auditory, because the percept (the chirp) is sound, but not speech. We have called the other way phonetic, because it has those characteristics we ordinarily perceive in consonants or vowels, and is manifest only when patterns of sound are being perceived as speech.

The second conclusion pertains to the role of silence. On the speech side of the duplex percept, silence had a large effect on the perception of formant transitions: when paired with silence, those transitions led to the perception of appropriate stop consonants; when paired with patterns containing no silence, they did not. On the nonspeech side, where the transitions were perceived as chirps, silence had no effect: though silence was paired with one transition and not with the other, most subjects correctly perceived that the corresponding chirps were the same when the transitions that produced them were the same, and different when the transitions were different. We suppose, therefore, that the importance of silence in the perception of stop consonants is owing to specifically phonetic (as distinguished from generally auditory) processes, and that the effect of silence in such cases is an instance of specifically phonetic (as distinguished from generally auditory) perception.

REFERENCES

- Bailey, P., & Summerfield, A. Q. Some observations on the perception of [s] + stop clusters. Journal of Experimental Psychology: Human Perception and Performance, in press.
- Darwin, C. J. The perception of speech. In E. C. Carterette & M. P. Friedman (Eds.), Handbook of perception, Vol. VII: Language and speech. New York: Academic Press, 1976, p. 175.
- Dorman, M. F., Raphael, L. J., & Liberman, A. M. Some experiments on the sound of silence in phonetic perception. Journal of the Acoustical Society of America, 1979, 65, 1518.
- Fitch, H. L., Halwes, T., Erickson, D. M., & Liberman, A. M. Perceptual equivalence of two acoustic cues for stop-consonant manner. Perception & Psychophysics, 1980, 27, 343-350.
- Isenberg, D. S., & Liberman, A. M. Speech and nonspeech percepts from the same sound. Journal of the Acoustical Society of America, 1978, 64, Suppl. 1, S20. (Abstract)
- Kuhl, P. K., & Miller, J. D. Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. Science, 1975, 190, 69.
- Liberman, A. M. Duplex perception and integration of cues: Evidence that speech is different from nonspeech and similar to language. In E. Fischer-Jørgensen, J. Rischel, & N. Thorsen (Eds.), Proceedings of the IXth International Congress of Phonetic Sciences, Vol. II. Copenhagen: University of Copenhagen, 1979, p. 468.

- Lieberman, A. M., Cooper, S., Shankweiler, D. P., & Studdert-Kennedy, M. Perception of the speech code. Psychological Review, 1967, 74, 431.
- Lieberman, A. M., & Studdert-Kennedy, M. Phonetic perception. In R. Held, H. W. Leibowitz, & H.-L. Teuber (Eds.), Handbook of sensory physiology, Vol. VIII: Perception. Berlin: Springer-Verlag, 1978, p. 143.
- Mattingly, I. G., Lieberman, A. M., Syrdal, A. K., & Halwes, T. Discrimination in speech and nonspeech modes. Cognitive Psychology, 1971, 2, 131.
- Miller, J. D. Perception of speech sounds in animals: Evidence for speech processing by mammalian auditory mechanisms. In T. H. Bullock (Ed.), Recognition of complex acoustic signals. Berlin: Dahlem Konferenzen Life Sciences Research Report 5, 1977, p. 49.
- Pastore, R. Contralateral cueing effects in the perception of aspirated stop consonants. Journal of the Acoustical Society of America, 1978, 64, Suppl. 1, S17. (Abstract)
- Rand, T. Dichotic release from masking for speech. Journal of the Acoustical Society of America, 1974, 55, 678.
- Repp, B. H., Lieberman, A. M., Eccardt, T., & Pesetsky, D. Perceptual integration of acoustic cues for stop, fricative, and affricate manner. Journal of Experimental Psychology: Human Perception and Performance, 1978, 4, 621.
- Studdert-Kennedy, M. Speech perception. In N. J. Lass (Ed.), Contemporary issues in experimental phonetics. New York: Academic Press, 1976.

FOOTNOTES

¹For reviews, see: Lieberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Darwin, 1976; Studdert-Kennedy, 1976; Lieberman & Studdert-Kennedy, 1978. For an opposing view—to wit, that the phenomena of speech perception can be accounted for by general auditory mechanisms—see, for example, Kuhl & Miller, 1975; Miller, 1977.

²Also see Isenberg & Lieberman, 1978; Pastore, 1978; Lieberman, 1979.

³In the case of stop-consonant manner, as with all phonetic features and segments, there are many acoustic cues. To change the percept by varying just one cue, it may, accordingly, be necessary to include some of the others—especially those that have considerable power—but to set them at neutral values. Thus, if we wish, by varying only silence, to interconvert convincingly between [spal] (or [stal]) and [sa], we will likely need an appropriate setting of at least one other cue—namely, a transition of the first formant that rises from a lower (but not too much lower) frequency. For that reason, we have shown such a transition in the schematic patterns said to be appropriate for 'normal binaural presentation'—that is, those of the upper row in the figure. As pointed out later in the text, and in Note 10, we found in working with the duplex percept that it was possible, and also to the advantage of the experiment, to use a straight first formant.

⁴The formant transitions were produced on the parallel resonance synthesizer of the Haskins Laboratories. The [p] transitions started at 924 and 1856 Hz and rose to 1077 and 2360 Hz; the [t] transitions started at 1306 and 2861 Hz and fell to 1077 and 2360 Hz. Their overall duration was 40 msec. When presented as the isolated transitions shown in the lower half of the

figure, they had 30 msec of steady-state formants appended to their trailing edges. The intensity of these appendages decreased gradually, complementing 30 msec of gradually increasing intensity in the corresponding formants of the 'base'.

5Our experience indicates that listeners will perceive the isolated transitions as more or less like speech depending on how many formants are presented. Given only the second- or third-formant transitions, few, if any, listeners will perceive a speechlike sound of any sort. Yet the phenomenon of duplex perception can be obtained when such single-formant transitions are isolated for dichotic presentation with the base.

6The 'fricative noise' of the base was produced from an s-like utterance (made by a human talker) in the following way. A 100-msec section of the noise was excerpted, low-passed at 5 kHz, and digitized. The amplitude envelope was tapered at either end, making the effective duration approximately 90 msec. There were energy peaks at about 3500 and 4500 Hz.

The remainder of the base--the three formants--was produced on the Haskins parallel resonance synthesizer. The first formant centered at 637 Hz and had a duration of 256 msec. The second and third formants centered at 1077 and 2360 Hz, respectively. They began 40 msec after the first formant, with gradually rising intensity over the first 30 msec, and ended simultaneously with it.

In the 'silence' condition, the fricative noise was separated from the formants by 100 msec, during which period there was little or no acoustic energy; in the 'no silence' condition, the corresponding interval was 20 msec.

7In the duplex percept the rising transition of the first formant, seen in the patterns at the top half of the figure, seems not necessary for perception of a stop consonant. Consequently, we straightened the first formant in the base component so that the isolated second- and third-formant transitions would be made to carry, not only the distinction between [p] and [t], but also much of the information about stop-consonant manner.

8Six members of the Speech Communication group at MIT, none of whom had served in the main part of the experiment, made phonetic identifications of the 'silence' and 'no silence' versions of the base, having been invited to use whatever phonetic categories they found most appropriate. The 'silence' version was identified 45% of the time as [s-a] (that is, as [s], followed by silence, followed by [a]) or as [s?a] (that is, [s], followed by glottal stop, followed by [a]), 28% of the time as [sa], and 27% of the time as [sta]. The 'no silence' version was identified 26% of the time as [s-a] or [s?a], 75% of the time as [sa]. The variability of these judgments was high, the standard error being about 15%.

9The base was always presented to the left ear and isolated transitions to the right ear. However, our experience indicates that duplex perception can be obtained with the reverse arrangement. If there is a difference in sensitivity between these two arrangements, it must be rather small; at all events, we have no evidence of it.

¹⁰The test tape was organized into 10 blocks of 8 trials each. Each block was a different permutation of the 8 types of trial. The inter-trial interval was 5 sec and the inter-block interval was 8 sec.

¹¹Each of 14 subjects who volunteered for the experiment was put through an initial screening procedure designed to determine whether he perceived the synthetic syllables appropriately and achieved the duplex percept. According to the outcome, four were excluded: two because they perceived some of the synthetic syllables as containing [l] or [w], one because he did not hear the 'chirp' as an isolable percept, and one because he was apparently not able to fuse the base and the isolated transition.

¹²Subjects responded by writing 'spa', 'sta', or 'sa' in the speech task and 'S' or 'D' in the chirp task. For the condition in which both tasks were carried out on each trial, subjects first wrote the 'speech' response, then the 'chirp' response.

¹³Of the three subjects who failed to discriminate the chirps accurately in the attend-chirp-only task, two were put through a further experiment to determine if they nevertheless could hear the chirp, especially when there was no silence in the base. This was done with stimuli in which the transitions and the base were sometimes presented dichotically and sometimes mixed electronically. In the former case, the subjects did hear chirps whether or not the silence was present; in the latter case, they did not perceive chirps at all.

¹⁴The improvement in performance in what would seem to be the more difficult task is owing to the performance of one subject; otherwise, discrimination of the chirps was less accurate when subjects had to make both speech and nonspeech judgments on each trial. In any case, an exact comparison is difficult because the difference between conditions is confounded with practice effects.

¹⁵A breakdown of the responses according to whether the transitions were the same or different revealed that, in the matter of errors, subjects were slightly more likely to judge 'same' when the stimuli were different than to judge 'different' when the stimuli were the same. If silence had affected perception of the chirps, the asymmetry in errors would have been in the opposite direction, since, for each pair of stimuli, silence was always paired with one but not the other.

¹⁶That the perception is duplex and not triplex--that the listener does not perceive the base as well as the fused pattern and the chirps--is important because it indicates that the listener is not processing the input in all possible ways, but only as speech and nonspeech. For the stimuli used in the present experiment, however, it can be objected that we should find it difficult to know whether or not the base (perceived as, let us say [s-a]) was present in the fused percepts [spa] or [sta]. We should remark, therefore, that in other research on duplex perception we have obtained clearer evidence for the disappearance of the base and hence for the essential duplexity of the result. Beginning with a base that was perceived very often as [ra], and adding a third-formant transition appropriate for [l], we got, as a result,

the percepts [la] and a chirp (D. S. Isenberg & A. M. Liberman, 1978, op. cit.).

It should be noted, too, that perception is not always duplex when an otherwise coherent syllabic pattern is divided into two parts which are then presented dichotically. Thus, to extend the example of [ra] vs. [la] described above, suppose that the first and second formants are presented as the base to one ear and all of the third formant (as opposed to only its transitional component) to the other. Listeners will then normally perceive only the (fused) syllable ([ra] or [la]) they would have perceived had the dichotically presented components been combined and presented in the normal way; they will not perceive anything resembling the sound of the third formant in isolation.

EVIDENCE FOR PHONETIC PROCESSING OF CUES TO PLACE OF ARTICULATION:
PERCEIVED MANNER AFFECTS PERCEIVED PLACE

Guy Carden,+ Andrea Levitt,++ Peter W. Jusczyk,+++ and Amanda Walley+++

Abstract. Using stimuli that could be labeled either as stops [b,d] or as fricatives [f,v,θ,ʒ], we found that, for a given acoustic stimulus, perceived place of articulation was dependent on perceived manner. This effect appeared for modified natural syllables with free-identification and forced-choice identification tasks, and for a synthetic transition continuum with a forced-choice identification task. Since perceived place could be changed by changing manner labels, with no change in the acoustic stimulus, it follows that the processing of the place feature depends on the value the listener assigns to the manner feature, rather than directly on any of the acoustic cues to manner. We interpret these results as evidence that the identification of place of articulation involves phonetic processing, and could not be purely auditory.

It is clear that a model for speech perception will include general-purpose auditory mechanisms at the input end of the system, and speech-specific phonetic or phonological mechanisms at the output end where segments are identified. The interesting theoretical question is, how far in the process can we go using mechanisms that are also used in non-speech processing? If we assume, for example, that the listener proceeds by first identifying features like place or manner, and then combining those features to identify segments, it is clear that the combination operation must include mechanisms that are specialized for speech processing ("phonetic mechanisms"). In principle, however, the mechanisms involved in processing the cues to the features might be purely auditory; and in fact it remains a matter of controversy whether phonetic mechanisms are needed to process the feature cues.

+Also Yale University.

++Also Wellesley College.

+++Dalhousie University.

Acknowledgment. This research was supported in part by NICHD Grant HD 01994 and BRS Grant 05596 to Haskins Laboratories, and in part by N.S.E.R.C. Grant A-0282 to Peter W. Jusczyk. We are grateful to Terry Halwes, Alvin Liberman, Rod McGuire, David Pisoni, Bruno Repp, Linda B. Smith, and Michael Studdert-Kennedy for useful advice and criticisms, though of course none of them necessarily agrees with us. We are also grateful to Janice Murray, Christopher Murphy, and Leon Serafim for assistance in running subjects.

[HASKINS LABORATORIES: Status Report on Speech Research SR-62 (1980)]

It is hard to get a clear test of this issue, since in most experiments the effects that seem to imply phonetic processing might imaginably be caused by complex auditory interactions; what we need is an experimental design that controls for auditory interactions. The best possible control is to hold the acoustic stimulus constant: If we can find a manipulation where the effect of interest can be triggered by a change in percept alone, then in principle no purely auditory explanation would be possible, and we would have to infer that phonetic processing was involved.

A series of experiments have succeeded in doing this, by constructing stimuli that could be perceived either as speech or as non-speech or as both simultaneously. Two different techniques have been used: the pure-tone synthesis used by Bailey, Summerfield, and Dorman (1977), Dorman (1979), and Best, Morrongiello, and Robson (1980), and the duplex perception used by Isenberg and Liberman (1978) and Liberman (1979). In each study, the authors investigated the familiar formant-transition cue to place of articulation. They manipulated the experimental conditions so that a given acoustic stimulus was perceived as speech in one condition and as non-speech in another. Finding differences in the way the stimuli were perceived in the speech and non-speech conditions (e.g., different discrimination functions), they inferred that the subjects were using special-purpose phonetic processing in the speech condition.

These experiments provided evidence for phonetic processing by showing that a given acoustic stimulus was processed differently in speech and non-speech conditions. It might be possible to construct a parallel argument for phonetic processing using only speech stimuli. Suppose we could find a case where a given acoustic stimulus is processed differently under two phonetic conditions. The stimulus involved cannot merely be ambiguous; we need to find cases where one phonetic dimension is dependent on a second, and then to manipulate the percept for that second dimension without changing the acoustic stimulus. The familiar dependencies among phonetic features offer a promising place to start: For example, Smith (1973), Miller (1977), and Eimas, Tartter, Miller, and Keuthen (1978) found evidence that in a number of cases the identification of place of articulation (e.g., labial vs. dental) was dependent on manner (e.g., stop vs. nasal, or stop vs. fricative). In these studies the manner dimension was cued acoustically: For example, Miller (1977) constructed synthetic voiced stop and nasal continua where the place feature was cued by variations in the transitions for F2 and F3 and the stop/nasal manner difference was cued by adding a nasal resonance and flattening the initial transition of F1 in the nasal stimuli. Miller found small but significant differences in the labial vs. dental boundary for the two continua. From our point of view, the interesting question is, does this boundary shift reflect phonetic processing, or only auditory interactions between the acoustic cues to manner and place?

The English stops and fricatives in the labial-to-dental¹ range look like a good choice for a study investigating whether changes in perceived manner will alter perception of place-of-articulation. Eimas et al. (1978) report a mutual but asymmetrical dependency in processing place and manner information for [ba, da, va, za], in which place is more dependent on manner than manner is on place. Moreover, we can find a parallel dependency in production if we select the right stops and fricatives to compare: For example, in the labial

range, the place of articulation for stops [p,b] is bilabial, while the place of articulation for the fricatives [f,v] is labiodental. Similarly, in the dental range, the stops [t,d] are alveolar, while the fricatives [θ,ð] are interdental. Thus, for English stops and fricatives, the dependencies observed for the perception of place and manner information seem to be mirrored, at least in part, by the presence of similar dependencies in production. If we look at the set [b,f,v,d,θ,ð],² a change in the manner feature will require a change in the place feature in articulation: The series runs from bilabial [b] to labiodental [f,v] to interdental [θ,ð] to alveolar [d]. This set has the further advantage that we can focus on a single cue to place, the formant transitions, which will be a sufficient cue to place both for the stops and for this set of fricatives. Harris (1958) gives evidence that the transitions are the primary cue distinguishing [f,v] from [θ,ð], and that there is little if any place information in the frication noise, contrasting with fricative comparisons like [s] vs. [ʃ], where the noise is the major place cue.

Let us assume then that the labial-to-dental changes in articulation in the stop-fricative series [b,f,v,d,θ,ð] correspond to formant transition changes that can be arranged along a single continuum. A simple auditory hypothesis predicts that the place boundary corresponds to a psycho-acoustic boundary. Given a single continuum, there would be no reason to expect that different psychoacoustic boundaries would appear depending on the manner. If we find different boundaries for the stops and the fricatives along the same continuum, and the difference cannot be attributed to an auditory interaction, then the difference between the boundaries must reflect phonetic processing. How could we show that the difference is not due to an auditory interaction? Recall that our first step was to establish a dependency, in this case a place boundary shift that depends upon the acoustically cued manner of the stimulus. Given this sort of dependency, what happens if we change only the perceived manner without changing the acoustic stimulus? Do we find the same dependency we found when the change in manner was cued acoustically? A negative result would not prove much, since it is natural to suppose that some feature dependencies may involve only auditory interactions, and in any case it would be difficult to prove that phonetic processing was NOT involved. But a positive result would be strong evidence: If, for example, we could shift a place boundary simply by changing perceived manner, phonetic processing must be involved; since the stimuli in the two conditions would be acoustically identical, in principle no purely auditory explanation would be possible.

Our objective in the following series of experiments was to investigate manner/place interactions in the set of stops and fricatives [b,f,v,d,θ,ð]. We need first to establish whether there is a manner/place dependency, which would be expected given the earlier work on feature dependencies with similar sets. If we find such a dependency, we need to change the perceived manner without changing the acoustic stimulus, and to test whether we find the same dependency we found when the manner change was acoustically cued. There are two main ways to perform such a manipulation: (1) We can construct stimuli that are ambiguous in manner, either by synthesis or by modifying naturally produced syllables. If we give the subjects a free-choice identification task, they will sometimes identify a given ambiguous stimulus as a stop and sometimes as a fricative. We can then observe whether this difference in perceived manner interacts with place identifications. We use this method in Experiment 1. (2) We can use a forced-choice identification task, and change

the instructions so that in one condition the subject is required to label a given set of stimuli as stops, and in the other condition he is required to label the same stimuli as fricatives. Again, we can observe whether the labeling interacts with place identifications. We use this method in Experiments 2 and 3.

EXPERIMENT 1

The purpose of Experiment 1 was to investigate the perception of the formant-transition cue to place of articulation in naturally-produced syllables, and, in particular, to see whether the perception of the place cue is affected by a change in perceived manner (stop vs. fricative). We therefore want to manipulate perceived manner without changing the naturally-produced formant transitions.

The primary cue to the stop/fricative manner contrast in CV syllables is the duration of the aperiodic noise that precedes the formant transitions. Fricatives have a relatively long noise (100-200 msec), a "frication;" stops have a short noise (typically a few msec), a "burst." Our preliminary work showed that, if one removes the burst or frication, the resulting truncated syllable is most often heard as a (voiced) stop, but is somewhat ambiguous between stop and fricative manner. Cutting off the frication from fricatives, therefore, seems to be a promising way of producing the ambiguous stimuli we need.

We therefore expect that the truncated fricatives will be perceived as ambiguous in manner, and we want to see whether the perceived place is different for the two perceived manners. We can exploit this ambiguity by giving the subjects a relatively natural free-choice identification task and simply examining the resulting pattern of perceived place for a given perceived manner. Our preliminary results showed that the truncated labiodental and interdental fricatives were most often identified as [f,v,b,θ,ʒ,d], so we used natural and truncated versions of these six syllables.

Method

Stimuli. Pretests showed that the labiodentals [f,v] gave reasonably consistent patterns of identification across speakers, but that there was substantial variation for the interdental [θ,ʒ]. We used these pretest results to select four adult male speakers who represented the observed range of variation. PN and PB were British; TH and DI were Americans.

For each speaker, we recorded 3 tokens of the CV syllables [fa], [va], [ba], [θa], [ʒa], and [da], using an Ampex AG500 tape recorder. Each token was digitized and low-pass filtered at 4.9 kHz using the Haskins Laboratories pulse code modulation (PCM) system (Cooper & Mattingly, 1969). We then used the PCM system to delete the bursts or frications from the digitized natural syllables, cutting at the point of zero or near-zero amplitude nearest to the end of the burst or frication.³ We will call the resulting truncated syllables [fa]-, [va]-, [ba]-, [θa]-, [ʒa]-, and [da]-.

Separate audio tapes were prepared for each speaker's tokens. Each tape consisted of 10 repetitions of each of the 3 tokens of the 6 natural

syllables, and 10 repetitions of the truncated versions of these 18 syllables. These 360 stimuli were arranged in a random order with a 3-second ISI and an additional 4 seconds after every 10th stimulus. The same randomization was used for all four tapes.

Subjects. Twelve college-age native speakers of English were paid for participating in two experimental sessions. None reported any history of speech or hearing disorder.

Procedure. The subjects were tested in small groups in a quiet room; the stimulus tapes were presented binaurally on an Ampex AG500 tape recorder through Grason-Stadler TDH 39-3002 headphones at a comfortable level. The subjects heard two tapes in the first session, and the other two in a second session a day or two later. The order of presentation was counterbalanced across subjects.

The subjects were told that the stimuli represented English consonant-vowel syllables, and that their job was to identify the initial consonant in each syllable by writing the appropriate letter on their answer sheet. Since [θ] and [ð] are represented by the same English letters 'th,' subjects were told to use 'th' for [θ] and 'dh' for [ð]; they practiced with minimal pairs like 'thigh' [θay] and 'thy' [ðay] until they reported that they understood the task. The instructions emphasized that all the stimuli began with an English consonant, and that if any stimulus sounded strange or ambiguous, the subject should write down the English consonant that sounded closest to what he heard. Subjects were, however, given the option of writing 'none' if they heard no consonant at all.

Results. Table 1 gives the results; for each syllable type, the responses are pooled for the 4 speakers and 12 subjects. Not surprisingly, the results for the truncated syllables are considerably noisier than those for the natural syllables. The complex pattern will be easier to interpret if we look at place and manner separately.

When a subject identifies a stimulus as beginning with a particular segment 'b' or 'f' or 'θ' or 'd,' he implicitly makes both a place judgment (labial vs. dental) and a manner judgment (stop vs. fricative). Recall that the purpose of the truncation was to create stimuli that were ambiguous between stop and fricative manner. This manipulation was reasonably successful: Overall, the truncated fricative syllables were identified as beginning with labial or dental stops 50% of the time, with labial or dental fricatives [f,v, θ,ð] 31% of the time, with other initial consonants 1% of the time, and with no initial consonant at all 17% of the time.

We now need to look for dependencies between the place and manner judgments implied by the segmental identifications. Table 2 gives a conditional probability chart for each truncated fricative, showing the proportion of labial and dental responses when that syllable was identified as a stop compared to the proportion when it was identified as a fricative. All four syllable types show an apparent dependency between perceived place and perceived manner: When a truncated fricative is identified as a fricative, the labial/dental proportion is essentially the same as for the original untruncated fricative; when the truncated fricative is identified as a stop,

Table 1

Percentages of Responses for Natural and Truncated Syllables

Experiment 1

Responses

Stimulus	Labial				Dental					Other
	p	b	f	v	t	d	θ	ð	ʃ	
[ba]		90	1	6				1	1	
[ba]-		59	6	3				1	31	
[fa]	1		80	3			14	1		
[fa]-		49	18	1			1	1	29	
[va]		1		79			3	16		
[va]-		68	4	20			1	2	5	
[da]						99				
[da]-		1				76	3	6	6	8
[θa]			16				80	3		
[θa]-		27	4	2		9	20	11	25	2
[ʃa]							18	81		
[ʃa]-		9		1		37	10	30	8	4

Table 2

Percentage of Labial and Dental Responses
Given Perceived Manner (Stop vs. Fricative)

Experiment 1

Responses

Stimulus	Manner	Labial	Dental
[fa]-	Stop	99	1
	Fricative	91	9
[va]-	Stop	100	-
	Fricative	88	12
[θa]-	Stop	75	25
	Fricative	16	84
[ʃa]-	Stop	20	80
	Fricative	3	97

the proportion of labial responses increases. The effect is relatively small for [fa]-, [va]-, and [ʋa]-; but very large for [θa]-, where the labial/dental proportion is almost reversed for the two perceived manners. We focus on the pattern for [θa]-, where the shift is significant not only for the pooled data, but for 10 of the 12 individual subjects ($p < .025$, using Chi-squared or an exact test as appropriate). This pattern holds up across subjects, speakers,⁴ and tokens, so that the pooled data in Table 2 represent a summation of cases where acoustically identical stimuli were perceived sometimes as stops and sometimes as fricatives, with more labial identifications when the syllable is perceived as a stop.⁵

Discussion

Experiment 1 supports earlier studies that found manner/place dependencies; it extends those studies in the limited sense that the dependency we found involves the interdentals [θ, ð].

The interesting feature of this dependency is that perceived place varies with changes in perceived manner even when the stimuli are acoustically identical. It follows that this dependency cannot be caused by any of the acoustic cues to manner or place, but must instead be caused by the subject's (necessarily phonetic) manner or place decision.

The contrast between [θa]- and the other truncated fricatives that we see in Table 2 suggests that there may well be two distinct mechanisms involved, one that applies to all the truncated fricatives and produces a significant but relatively small effect, and one that applies only or most strongly to [θa]-, and produces a more substantial shift in the same direction. In the following experiments, we will focus on the stronger effect that we found with [θa]-; we will discuss the other syllables briefly in the General Discussion. Because the task in Experiment 1 was a free identification, we have no evidence whether the place/manner dependency is mutual or asymmetric, with one feature controlling the other. Previous work on place/manner dependencies (Miller, 1977; Eimas et al., 1978) suggests that we are likely to find a mutual but asymmetric dependency, with place more dependent on manner than vice-versa. We can test for such a directionality by manipulating perceived manner directly with a forced-choice task.

EXPERIMENT 2

The most important result of Experiment 1 was the observation that there was a dependency between perceived place of articulation and perceived manner. The purpose of Experiment 2 is to replicate this result under more tightly controlled conditions, and to test the directionality of the dependency.

In Experiment 1 we used a free-identification task with a stimulus that was ambiguous in manner; in Experiment 2 we used one of the same stimuli (a token of [θa]-), but replaced the free-identification task with a forced-choice task: In one condition, the subjects had to choose between 'f' and 'th,' and so had to label the stimuli as fricatives; in the other condition, they had to choose between 'b' and 'd,' and so had to label the stimuli as

stops. To make it easy to hear the [θa]- with the required manner, we recorded the stimuli against a constant background of white noise.

Method

Stimuli. This experiment requires a token of [θa]- that is reasonably ambiguous in manner. We therefore selected the most promising token from our pretest data, one of the tokens from TH used in Experiment 1. As controls we used the corresponding untruncated [θa], and one token each of [ba], [da], [fa], and the corresponding [fa]-. All of these stimuli were prepared in the way described in Experiment 1.

We prepared two audio tapes. The "stop tape" contained natural [ba] and [da] and truncated [fa]- and [θa]-; the "fricative tape" contained natural [fa] and [θa] and the corresponding truncated [fa]- and [θa]-. We recorded a constant background of white noise at a signal/noise ratio of approximately 6 dB relative to the peak amplitudes of the vowels. Each tape contained 15 repetitions of each of the four stimuli, arranged in a random order with a 3-second ISI and an additional 3 seconds after every 10th stimulus.

Subjects. The same 12 subjects participated in Experiment 2 after the second experimental session of Experiment 1.

Procedure. The test setting was the same as in Experiment 1. The presentation order of the stop and fricative tapes was counterbalanced across subjects.

The subjects were told that they would be hearing English consonant-vowel syllables recorded in noise. For the stop tape, they were told to decide whether the initial consonant was 'b' or 'd,' and to indicate their choice by circling 'b' or 'd' on their answer sheet. The instructions for the fricative tape were the same, with the choice 'f' or 'th.' The subjects were told that they should guess if a stimulus was hard to identify.

Results. Table 3 gives the proportion of labial [b,f] and dental [d,θ] responses for the [θa]- stimulus in the two conditions.

Table 3

Percentage of Labial and Dental Responses
Given Perceived Manner (Stop vs. Fricative)

Experiment 2		Responses	
Stimulus	Manner	Labial	Dental
[θa]-	Stop	94	6
	Fricative	46	54

In the fricative condition, performance was essentially at chance, with [θa]-labeled as dental 'tha' 54% of the time. In the stop condition, the acoustically identical stimulus was labeled as labial 'ba' 94% of the time. When we examine the individual subjects, the numbers are small enough so that only 6 of the 12 show a significant difference, but all 12 shift in the same direction, with more labial responses in the stop condition. This 12-to-0 directionality is itself significant by a sign test ($p < .01$).

Discussion

Experiment 2 replicated the main result of Experiment 1 using a different task: When we controlled perceived manner by specifying the labels used in a forced-choice task, a change in labels led to a change in the pattern of identifications. The same acoustic stimulus was identified as labial 'ba' 94% of the time when perceived as a stop, but when it was perceived as a fricative, it was identified as dental 'θa' 54% of the time.

Experiment 2 also provides some evidence about the direction of the dependency: Since we manipulated perceived manner directly, the results show that perceived place is dependent on perceived manner. Of course it is possible that there may also be a dependency in the opposite direction, with manner dependent on place, so that there would be a mutual place/manner dependency, as suggested by Eimas et al. (1978).

What is crucial to our main argument is that the place shift could be cued by a change in instructions (and control stimuli) alone: Since the acoustic stimulus was the same [θa]- in both conditions, the shift must be triggered by the phonetic manner percept, rather than directly by any of the acoustic cues to manner.

EXPERIMENT 3

The primary purpose of Experiment 3 was to explore further the dependency between perceived manner and perceived place that was implied by the first two experiments. We used synthetic stimuli and both identification and discrimination tasks; as in Experiment 2, we manipulated perceived manner by changing the labels for the stimuli in a forced-choice identification task. The conditions differed from Experiment 2 in that there was no noise background and the stimuli were not constructed to be ambiguous in manner.

Method

Subjects. The subjects were 20 college-age native speakers of English from Dalhousie University. Each subject was paid \$4 for participating in two experimental sessions on successive days. None of them had had substantial previous experience with synthetic speech.

Stimuli. The stimuli were constructed using the OVE IIIc serial synthesizer. We synthesized two 9-item continua: a [ba] to [da] stop continuum, and a [fa] to [θa] fricative continuum using the same set of formant transitions, but with a uniform neutral frication at the beginning of each syllable.

Even though there is evidence (Harris, 1958) that the transitions are the primary cue to place for a choice between [fa] and [θa], with relatively little place information in the frication, we still need to take precautions to produce a frication that is neutral between [f] and [θ]. To do this, we used information from the FOVE (Synthesis by Rule) program at Haskins and chose OVE parameters that were midway between the FOVE parameters for [f] frication and [θ] frication. The resulting frication had pole values of 1233 Hz and 2540 Hz and a zero value at a level 5 dB below the steady-state part of the vowel. In addition, an aperiodic noise source excited F1, F2, and F3 at 200, 1404, and 2749 Hz. The frication was 126 msec long.

To produce a natural-sounding vocalic portion, and to center our transition continuum on the labiodental-to-interdental region, we began by recording tokens of [fa] and [θa] spoken by one of the authors (PWJ). We used these natural tokens as the basis for a copy synthesis: The waveform was sampled at a rate of 10 kHz and bandpassed at 90 Hz to 4.9 kHz, providing input for a ubiquitous spectrum analyzer taking 128 samples per frequency span (approximately 4.9 kHz bandwidth), repeated every 12.8 msec. The spectrum data were wideband filtered (240 Hz), and we used a peak-extraction routine to provide values for the parameters for the OVE IIIc synthesizer.

Given parameters for the vocalic portion of [fa] and [θa] from the copy synthesis, we constructed the rest of the transition continuum as follows: First, we took the copy synthesis versions of [fa] and [θa] as the endpoints of a 5-item continuum generated by interpolating the starting frequency values for F2 (in steps of about 43 Hz) and F3 (in steps of about 95 Hz). Second, we extended the continuum two steps beyond each of the previous endpoints; thus the copy synthesis versions of natural [fa] and [θa] correspond to stimuli 3 and 7 in the final continuum. Table 4 gives the starting frequency values for the transitions for the whole continuum. While the starting frequencies varied, all the stimuli reached steady-state values for F2 of 1198 Hz after a 42 msec transition and for F3 of 2396 Hz after a 54 msec transition. All stimuli had the same F1 and fundamental frequency contours: F1 began at 479 Hz and rose to a steady-state value of 777 Hz after a 54 msec transition; the fundamental began at 138 Hz, peaked 60 msec later at 144 Hz, and then gradually declined to 118 Hz. The whole vocalic portion, transitions plus steady-state vowel, was 192 msec long.

Without the frication, listeners identified these stimuli as syllables beginning with voiced stops, ranging from 'ba' (stimulus 1) to 'da' (stimulus 9). When the frication was attached immediately before the transitions (giving a stimulus 318 msec long), listeners identified the new stimuli as syllables beginning with voiceless fricatives, ranging from 'fa' (stimulus 1) to 'θa' (stimulus 9).

We prepared identification and "oddity" discrimination tapes for both continua. The identification tapes had 15 tokens of each of the 9 stimuli

Table 4

Starting Formant Frequencies for F2 and F3

Experiment 3

Stimulus Number	F2	F3
1	1107	2150
2	1147	2245
3	1198	2345
4	1242	2431
5	1288	2539
6	1335	2632
7	1374	2729
8	1425	2809
9	1467	2912

(steady-state F2 = 1198 Hz, F3 = 2396 Hz)

Table 5
Design of Experiment 3

	Day 1	Day 2	Day 2
Group I (5 Ss)			
Identification	[ba]-[da]	[fa]-[θa]	[fa]-[θa] relabeled as [ba]-[da]
Discrimination	[ba]-[da]	[fa]-[θa]	[fa]-[θa]
Group II (5 Ss)			
Identification	[ba]-[da]	[fa]-[θa]	[ba]-[da] relabeled as [fa]-[θa]
Discrimination	[ba]-[da]	[fa]-[θa]	[ba]-[da]
Group III (5 Ss)			
Identification	[fa]-[θa]	[ba]-[da]	[ba]-[da] relabeled as [fa]-[θa]
Discrimination	[fa]-[θa]	[ba]-[da]	[ba]-[da]
Group IV (5 Ss)			
Identification	[fa]-[θa]	[ba]-[da]	[fa]-[θa] relabeled as [ba]-[da]
Discrimination	[fa]-[θa]	[ba]-[da]	[fa]-[θa]

arranged in a random order with a 3-second ISI. In the discrimination tapes, stimuli were presented in groups of three, with each triad composed of two tokens of one stimulus and one token of a second stimulus three steps away on the continuum, for example, Stimulus 1 vs. Stimulus 4. For each comparison there are six possible arrangements (AAB, ABA, BAA, BBA, BAB, ABB); there are six possible three-step comparisons (1-4, 2-5, 3-6, 4-7, 5-8, 6-9); thus we have a total of 36 triads. The stimulus tapes had two tokens of each triad, giving a total of 72 stimuli, arranged in a random order with a 1-second interval between elements of a triad and a 3-second interval between triads.

Procedure. Subjects heard the stimuli over Koss Pro 4AAA headphones. Using a sound level meter (General Radio model 1565-A), we adjusted the volume so that the stimuli were played at a level of approximately 72 dB (A) SPL. Subjects listened to the stop and fricative tapes on successive days, with the order counterbalanced across subjects. On each day, they did the identification task before the discrimination task. Following their day-2 identification and discrimination tasks, the subjects did a relabeling task where they were required to identify stimuli from one continuum using labels appropriate to the other continuum. Thus half the subjects heard the stop tape as a forced choice 'f' vs. 'θ,' and the other half heard the fricative tape as a forced-choice 'b' vs. 'd.' After this relabeling task, the subjects repeated the discrimination task for the stimuli they had just relabeled. Table 5 gives the complete design of the experiment.

Instructions

For the identification tapes, subjects listening to the stop tape were told that the stimuli represented 'ba' or 'da,' and that they should write 'b' or 'd' on their answer sheets accordingly. Similarly, subjects listening to the fricative tape were told that the stimuli represented 'fa' or 'tha,' and that they should write 'f' or 't' accordingly. For the discrimination task, subjects were told that they would be hearing stimuli in groups of three, and that they were to decide which stimulus in each group differed from the other two; they indicated their choice by circling the appropriate letter (A, B, C) on a prepared answer sheet. For the relabeling task, subjects were told that we wanted them to label some of the stimuli they had already heard with new labels: subjects relabeling the stop tape were told to identify the stimuli as 'fa' or 'tha,' and to write 'f' or 't' accordingly, and vice versa for the subjects who relabeled the fricative tape. Subjects were told that they might find the task difficult, but that they should do the best they could. For the discrimination task that followed the relabeling task, the subjects were told that they were hearing the same stimuli again, and that they could label the stimuli in any way that helped them to do the task, using the new labels, the original appropriate labels, both, or neither.

Results

Identification. Figure 1 gives the labeling functions for the stop and fricative continua, plotting the mean number of times, across all subjects, that each stimulus was identified as labial ('f' for the fricative continuum, 'b' for the stop continuum). It is clear from Figure 1 that the cross-over

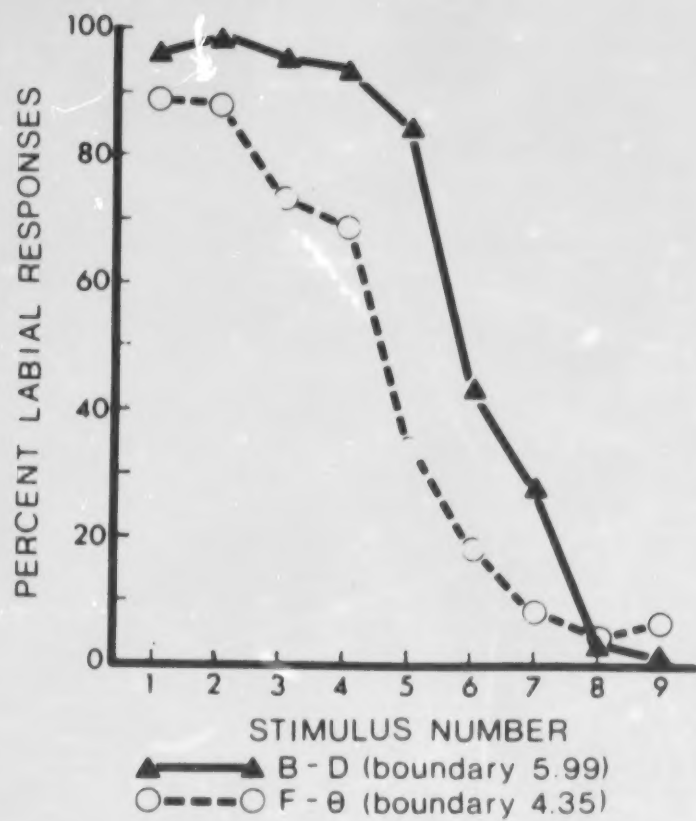


Figure 1

for the fricative function is considerably further forward (further toward the labial end) than that of the stop function. This in itself suggests that the locus of the place-of-articulation boundary is different for the two series. We estimated stop and fricative boundaries for each subject by fitting a normal ogive function to his or her identification data (Kling & Riggs, 1971); the two left-hand columns in Table 6 list the individual boundaries. Note that, for all twenty subjects, the fricative boundary is further forward than the stop boundary. The mean boundaries were 4.35 for the fricative continuum and 5.99 for the stop continuum; this difference is significant well beyond the $p < .01$ level, $t(19) = 13.04$.

Discrimination. Figure 2 gives the mean three-step discrimination functions for the stops and fricatives, along with predicted discrimination functions calculated from the identification data (Liberman, Harris, Hoffman, & Griffith, 1957; Pollack & Pisoni, 1971). The peak for the fricatives seems to be somewhat forward of the peak for the stops, consistent with the results for the identification task. To test this, the data were submitted to an ANOVA of a 2 (series: stop vs. fricative) $\times$ 2 (obtained vs. predicted) $\times$ 6 (stimulus comparison, e.g., 1-4, 2-5 etc.) repeated measures design. This analysis showed significant main effects for both series, $F(1,19) = 20.15$, $p < .01$, and stimulus comparison, $F(1,19) = 23.25$, $p < .01$. The first shows that the stops are significantly more discriminable than the fricatives, while the second shows that both discrimination functions have significant peaks. In addition, the interaction between series and stimulus comparison was also significant, $F(5,95) = 23.75$, $p < .01$, showing that the peaks occurred in different locations for the two series. There was no evidence of a significant main effect for obtained vs. predicted, $F(1,19) = 2.6$, and no interactions with this factor approached significance. This shows good agreement between the identification and discrimination data, on the assumption that perception of place is categorical in both stops and fricatives.

Relabeling. Subjects were instructed to identify stimuli from one series using labels appropriate to the other. Ten subjects relabeled the [ba]-[da] stimuli from the stop series as fricatives 'fa,' 'θa'; the other ten relabeled the [fa]-[θa] fricative stimuli as stops. Figure 3 shows the mean identification functions from the relabeling task for each group, together with the mean identification functions from the same subjects' original (appropriate) labeling of the same stimuli. It is clear that the effect of the instructions in the relabeling task was to shift the identification function in the direction of the new labels. For example, when the stops were relabeled as fricatives, the new identification function was shifted forward, toward the location of the original fricative identification function. As before, we calculated place boundaries for the individual subjects from ogives; the individual boundaries are listed in the right-hand column of Table 6. When we compare the boundaries from the relabeling task with those from the original identification task where the same stimuli were identified using the acoustically appropriate labels, we see that for all twenty subjects the relabeling boundary shifts from its original location toward the boundary corresponding to the new labels. Matched-pair t -tests confirm that the size of the shift is significant both for fricatives relabeled as stops, $t(9) = 4.623$, $p < .01$, and for stops relabeled as fricatives, $t(9) = 4.988$, $p < .01$.

Table 6

Individual Subject's Category Boundaries for the Identification
and Relabeling Conditions in Experiment III

Subject Number	Identification	Relabeling	
	b-d	f-θ	f-θ→b-d
1	5.19	4.18	4.27
4	6.53	4.56	5.63
5	4.59	3.04	4.49
7	5.80	3.45	5.19
11	6.27	3.53	5.39
12	6.62	5.16	5.22
13	6.31	4.20	5.87
14	6.19	4.46	4.73
15	5.43	3.90	5.00
16	5.45	4.40	5.15
$\bar{X}$ (S.D.)	5.84 (.66)	4.09 (.62)	5.09 (.49)

Subject Number	b-d	f-θ	b-d→f-θ
2	5.57	3.96	5.10
3	6.46	4.66	3.57
6	6.24	4.68	5.20
8	5.22	3.39	1.91
9	5.14	4.62	3.55
10	5.87	4.32	4.81
17	6.07	4.30	4.36
18	7.79	7.26	4.38
19	6.03	3.63	5.71
20	7.00	5.22	5.22
$\bar{X}$ (S.D.)	6.14 (.81)	4.60 (1.08)	4.38 (1.12)
Grand $\bar{X}$ (S.D.)	5.99 (.73)	4.35 (.90)	

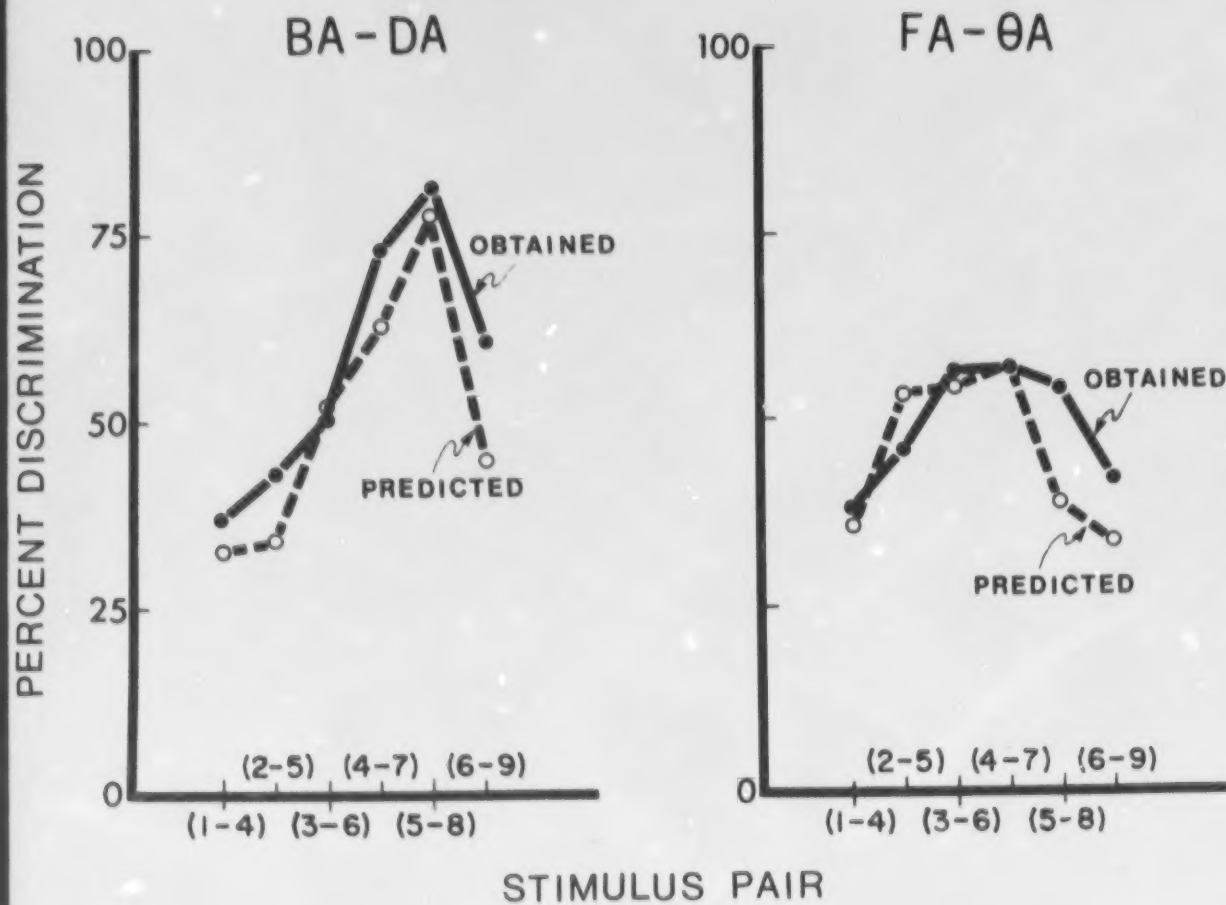


Figure 2

Relabeling Identification

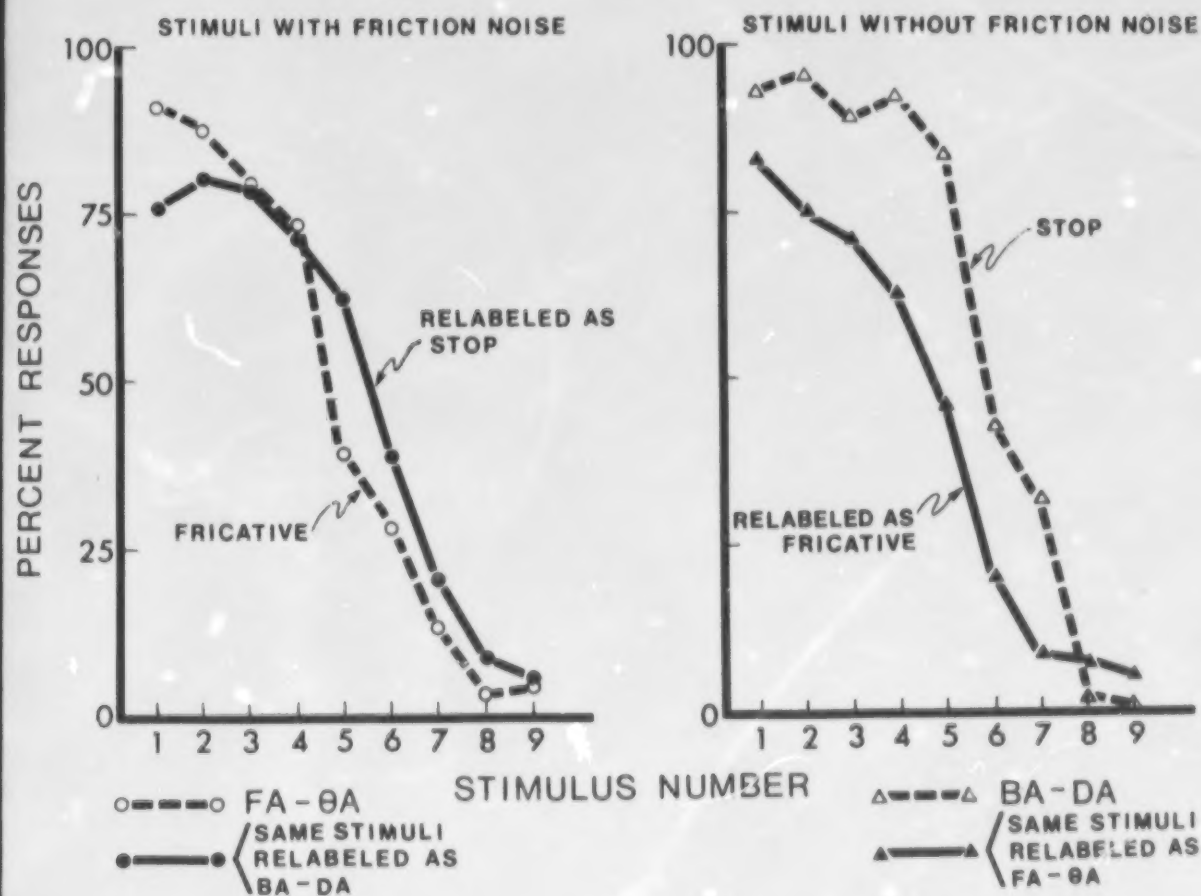


Figure 3

It is clear that the change in instructions, and so presumably a change in the manner percept, shifted the boundary in the predicted direction. In the original identification task, the stop and fricative conditions differed both in instructions and in acoustic cues to manner. Does the change of instructions by itself produce as big a boundary shift as the one we found when we compared the place boundaries for acoustically cued stops and fricatives? Here we found an interesting asymmetry: There was no significant difference between the original fricative boundary and the boundary we found when the stops were relabeled as fricatives, $t(9) = 0.492$; that is, a change in the instructions alone seemed to make the acoustic stops shift all the way to the fricative boundary. But there was a significant difference between the original stop boundary and the boundary we found when the fricatives were relabeled as stops, $t(9) = 5.207$, $p < .01$. In other words, the fricatives relabeled as stops shifted in the direction of the stop boundary, but less than all the way.

Post-relabeling discrimination. Figure 4 gives the mean discrimination functions for the ten subjects tested on each continuum after the relabeling task, compared with the original mean discrimination functions for the same subjects on the same stimuli. The peaks of the discrimination functions seem to have shifted slightly in the direction of the new labels, but the striking result is the improved discriminability of 4-7, 5-8, and 6-9 when the fricatives were relabeled as stops, and the improved discriminability of 3-6 and the reduced discriminability of 5-8 and 6-9 when the stops were relabeled as fricatives. The tendency is for discriminability to improve around the boundary that is appropriate to the new labels, and to decline when the comparison crossed the old boundary but not the new. For example, 3-6 crosses the fricative boundary but not the stop boundary for the subjects who relabeled the stops as fricatives, and so discriminability improves, while 5-8 crossed the stop boundary but not the fricative boundary, and so discriminability declines. An ANOVA of a 2 (series: stop vs. fricative) $\times$ 2 (original discrimination vs. post-labeling discrimination) $\times$ 6 (stimulus comparison) design showed a significant main effect of stimulus comparison, $F(5,90) = 17.32$, $p < .01$, and a significant three-way interaction among series, original vs. post-relabeling, and stimulus comparison, $F(5,90) = 4.75$, $p < .01$. These results therefore show that the relabeling instructions had a significant effect on discrimination as well as identification, even though the subjects were not instructed to use the new labels for the discrimination task.

Discussion

The results from the first identification tasks show that a synthetic [ba]-[da] transition continuum has a different place boundary from the same transition continuum preceded by a [f/θ] synthetic frication and labeled as fricatives 'fa,' 'θa.' The results of the first discrimination task show that the corresponding discrimination functions shift in the same way. It is not surprising to find such a place/manner dependency, given the results of earlier studies like Miller (1977); our results extend these earlier studies to the limited extent that the dependency involves [θ], and that we found dependencies in discrimination as well as identification.

Post - Relabeling Discrimination

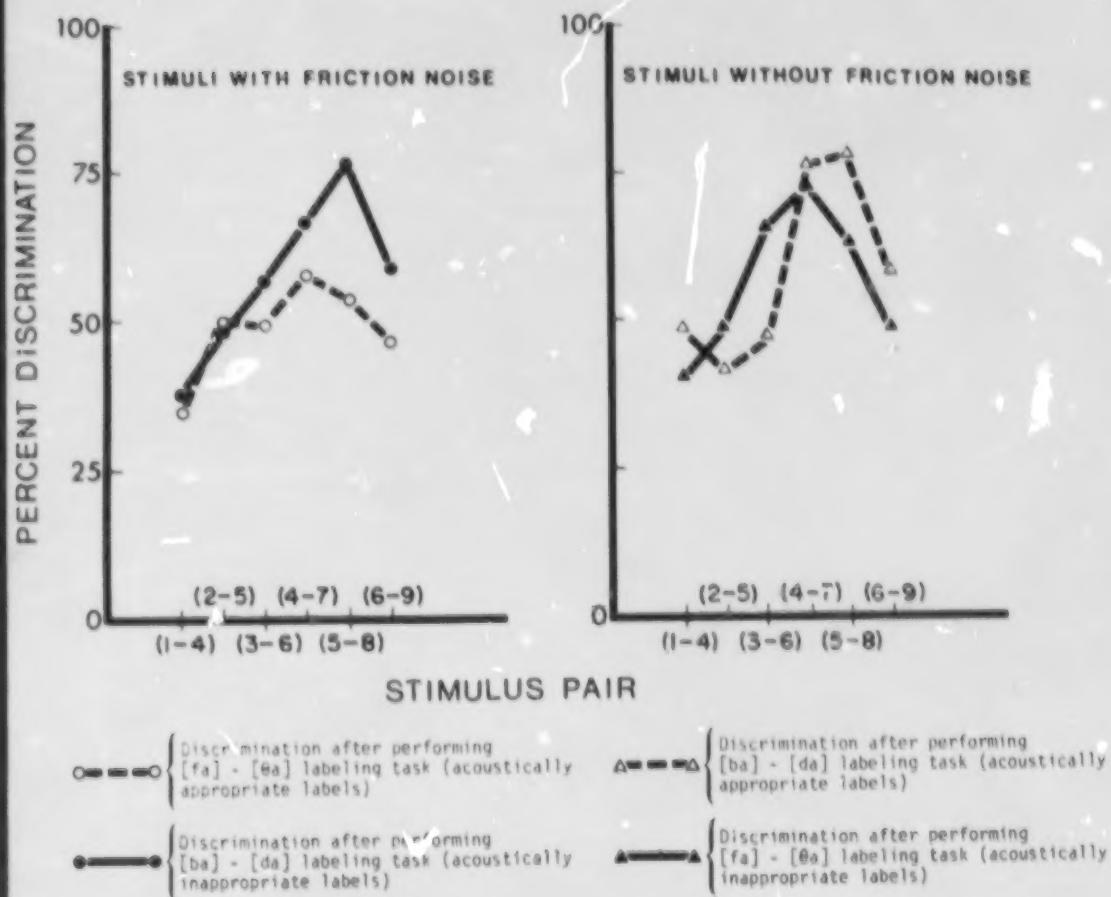


Figure 4

By themselves, the original identification and discrimination tasks do not bear on the question of phonetic processing: The differences we observe between the stop and fricative series might be caused by a dependency in phonetic processing, but they might equally well be caused by an auditory interaction between the frication and the transitions, by a trading relation between the place cue in the transitions and a place cue in the frication (if we hypothesize that our synthetic frication was a dental cue), or by a combination of all three.

We get evidence bearing on phonetic processing only when we compare the results of the original identification and discrimination tasks with the results of the relabeling task and the post-relabeling discrimination. This comparison shows that both identification and discrimination functions can be changed simply by changing the instructions. Whether or not auditory interactions played a role in producing the different place boundaries for the acoustically-cued stop and fricative series, they could not have produced the contrasts in Figures 3 and 4, where we found different identification and discrimination functions for the same acoustic stimuli, depending only on the labeling instructions given to the subject.

Experiment 3 therefore confirms the main result of Experiments 1 and 2: Perceived place is dependent on labeling, and so presumably on perceived manner. Experiment 3 extends our earlier result by showing that a change in manner labeling can affect discrimination as well as identification.⁶

GENERAL DISCUSSION

In each of the preceding experiments we manipulated perceived manner without changing the acoustic stimulus.⁷ In Experiment 1, we constructed a stimulus that was ambiguous in manner and observed a dependency between the perceived place and manner features that we deduced from segment identifications. In Experiments 2 and 3, we manipulated perceived manner directly by changing the instructions in a forced-choice task. Again, when the perceived manner changed, perceived place also changed. Note that this result holds for both forced-choice and free identification tasks, and for both synthetic and (modified) natural stimuli.

We conclude that the perception of place is dependent on manner, consistent with earlier studies like Eimas et al. (1978). Our result differs from the results of these earlier studies, however, in that the shift in perceived place can occur with acoustically identical stimuli. It follows that perceived place is dependent on the (necessarily phonetic) manner percept itself, rather than directly on any of the acoustic cues to manner. This result is consistent with models like the response-conditional model of Smith (1973), the prototype model of Oden and Massaro (1978), or the models implied by Miller (1977) and Eimas et al. (1978). It is inconsistent with any model that handles the perception of place, or dependencies between place and manner, entirely by acoustic or auditory mechanisms.

The Two-Boundary Hypothesis

Our results show that perceived place is dependent on perceived manner, but they do not provide direct evidence about what produces this dependency. It seems clear that the perception of one or more of the cues to place is affected by perceived manner, but there are several imaginable cues that might be involved.

The most natural interpretation, and the one we at present favor, is that perceived manner affects the perception of the formant-transition cue. In Experiment 3, the synthetic stimuli in a given continuum were identical except for the formant transitions, and it is therefore natural to assume that the place judgments were cued solely or primarily by the transitions. If this is so, the results of Experiment 3 would be accounted for by what we could call the "two-boundary hypothesis": Suppose that the listener can process an incoming transition using either of two boundaries. When he has identified the manner as stop, he uses one boundary; when he has identified the manner as fricative, he uses the other, which is slightly further forward (toward the labial end of the continuum), as implied by the boundaries listed in Table 5.

How would we apply this two-boundary hypothesis to the results of Experiments 1 and 2, where we were dealing with truncated natural syllables rather than with a synthetic continuum? Let us represent a transition continuum like the synthetic one schematically by a line (Figure 5); we could think of this line as representing a perceptual transition space, or, more concretely, as representing the range of formant-transition starting frequencies for some given vowel. Recall that the naturally-produced stops [b,d] differ in place of articulation from the corresponding fricatives [f,v,θ,ð]. The endpoints of the line in Figure 5 correspond to the transitions produced by bilabials [b] and alveolars [d]. Labiodentals [f,v] and interdentals [θ,ð] have intermediate places of articulation, and therefore presumably produce intermediate transitions. The fricative boundary must fall between the points corresponding to the natural labiodental and interdental transitions; Experiment 3 shows that the stop boundary must fall further back. We hypothesize that the stop boundary falls on or behind the point corresponding to the natural interdental articulation, as sketched in Figure 5. This hypothesis will account for the data of Experiments 1 and 2 as follows: When the ambiguous [θa]- stimulus is identified as a fricative in Experiment 1, or when there is a forced-choice between fricative responses 'f' and 'θ' in Experiment 2, the subject uses the fricative boundary to evaluate the transition cue, and so gets the same labial/dental distribution that he got for untruncated [θa]. When [θa]- is identified as a stop, or when the choice is between the stops 'b' and 'd,' the subject uses the stop boundary. Since the transitions produced by the [θa] articulation fall, on the average, slightly on the labial side of the stop boundary, the subject identifies the syllable mostly as labial 'ba' rather than dental 'da.'⁸

Alternative Hypotheses

While we feel that this two-boundary hypothesis gives the most promising account of our data, our results are also consistent with the hypothesis that there is a single perceptual boundary for the transition cue, and that the

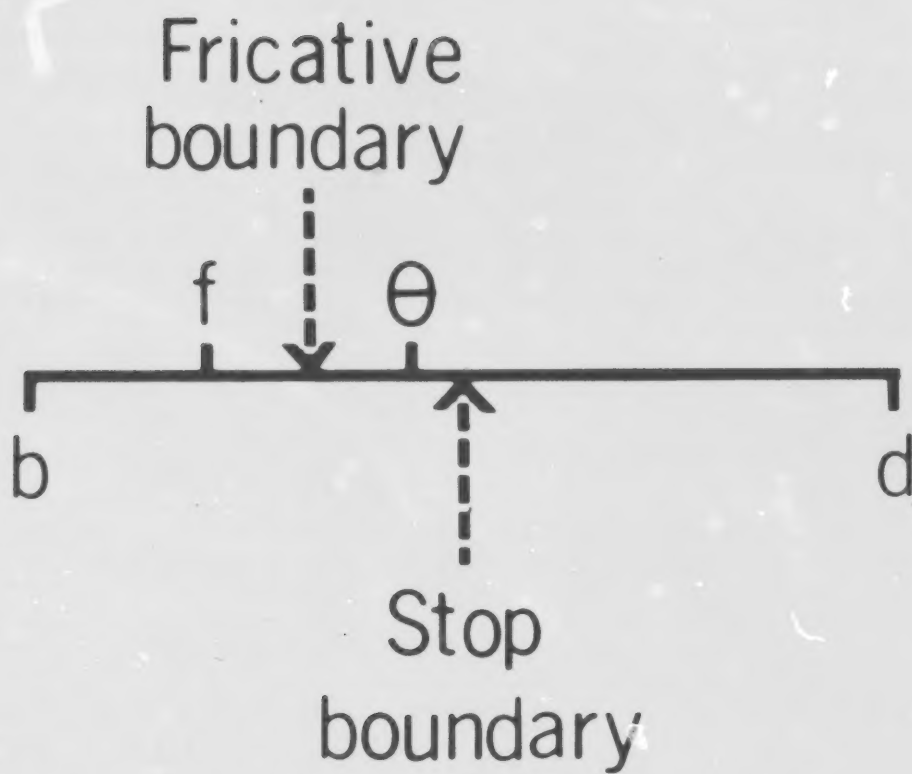


Figure 5

shift in place identifications (and so the shift in the apparent boundary in the synthetic transition continuum) results from a trading relation between the transition cue and some other place cue, X. In such an analysis, it would be the perception of cue X that was dependent on perceived manner and so was crucially phonetic. In other words, our results show that the perception of some place cue is sensitive to perceived manner, but they do not show that the relevant cue is necessarily the transition cue itself.

While we have not done any direct investigation of alternative cues, at least two candidates for cue X are imaginable. (1) Our stops consistently lacked bursts; [b] bursts are typically smaller in amplitude than [d] bursts. Lack of burst might therefore be a labial cue, and it would be natural for such a cue to be relevant only when the perceived manner was stop. (2) The truncated fricatives in Experiments 1 and 2, and the stops relabeled as fricatives in Experiment 3, all lacked acoustic frications. [θ, ð], for some speakers, have frications the length of stop bursts, markedly shorter than [f, v] or other fricatives. Length of frication might therefore be a place cue for the choice [f, v] vs. [θ, ð], with short or zero frication being a cue for dental. Again, it would be reasonable for such a cue to be relevant only when the perceived manner was fricative.

Either of these putative cues could imaginably produce the observed place/manner dependency: When the truncated fricatives and synthetic stops were perceived as stops, the labial cue from lack-of-burst would increase the number of labial identifications relative to the same transitions perceived as fricatives. When the truncated fricatives and relabeled synthetic stops were perceived as fricatives, the dental cue from short-frication would increase the number of dental identifications relative to the same transitions perceived as stops. The short-frication cue is less attractive, because it would not account directly for the fact that we also got more dental identifications when the fricative percept was cued by a full-length frication: We would have to assume that this shift was caused independently by place cues in the frication itself, so that our synthetic frication was a dental cue. We emphasize again that our results show that whatever cue is involved must be sensitive to perceived manner; the argument for phonetic processing cannot be escaped by adopting one of the alternative cues.

The Place/Manner Dependencies with [f], [v], and [θ]

Our experiments and discussion have focused on the place/manner dependency with [θa], but we must also consider how to account for the smaller dependencies we saw for the other truncated fricatives in Experiment 1. We have little to say about [fa]- and [va]-; we suspect that the shift may be caused by a trading relation between the transition cue and some other place cue introduced by the truncation operation, perhaps the lack-of-burst or the short-frication cue discussed above. But we have no evidence bearing on this issue. The case of the voiced interdental [ða]-, however, requires some discussion: Taken by itself, the two-boundary hypothesis implies that the voiced and voiceless interdentals should act alike. Recall, however, that while the voiced [ða]- showed a significant place/manner dependency, the size of the effect was much smaller for [ða]- than for [θa]-: When [ða]- was identified as a stop, the identifications were only 20% labial 'ba,' contrast-

ing with 75% 'ba' identifications for [θa]-. We need to see if we can explain this voiced/voiceless contrast within the two-boundary hypothesis. One possibility is that the perception of the transition cue might be dependent on voicing as well as manner, as suggested by the results of Miller (1977) and Eimas et al. (1978). A second possibility, perhaps sufficient in itself to explain our data, is that [β] may be articulated slightly further back than [θ], so that the transitions in [βa]- fall slightly on the dental side of the stop boundary. While textbooks do not report any difference in position of articulation between [θ] and [β], an informal survey of phonetically-trained native speakers of English showed general agreement that [β] is articulated slightly further back than [θ], consistent with the perceptual contrast we found in Experiment 1.

If our informants are correct in introspecting that [β] is articulated further back than [θ], then we can make an indirect argument for the two-boundary hypothesis: There is no reason to suppose that any of the other candidates for cue X ought to be sensitive either to voicing or to small differences in place of articulation. To explain the contrast between [θa]- and [βa]-, we would need to postulate a voicing/place dependency, and further, to limit that dependency to interdentals, unless the lack of a contrast between [fa]- and [va]- can be explained as a ceiling effect. This is not impossible, but the two-boundary hypothesis has the advantage of providing a direct explanation: If we make a change in position of articulation, that will produce different transitions and so potentially a different output from the mechanisms processing the transition cue.

Conclusion

We find the alternative hypotheses implausible, but our primary reason for preferring the two-boundary hypothesis is that it offers a possible explanation for the existence of this place/manner dependency: A boundary "efficiently" located between bilabial transitions and alveolar transitions would not be expected to coincide with a boundary efficiently located between labiodental and interdental transitions. On this view, the difference in place of articulation in production is the reason for the different boundaries in perception. This rationalization could apply only if the transition cue was crucial; if lack of burst or length of frication were crucial, then we would expect the perceptual dependency to be independent of the articulatory place/manner dependency we have in English. The validity of this production-based explanation for the perceptual dependency can therefore be tested directly by investigating the perceptual dependency in languages that lack the English articulatory dependency.9

While we have not been able to reach a firm conclusion about which place cue is dependent on the manner percept, we wish to emphasize that our major theoretical conclusion is independent of which cue is involved:

The results of our three experiments agree in showing that the processing of the place feature will have to be sensitive to perceived manner, rather than directly to any combination of the acoustic cues to manner. Because the shift occurs even when the stop and fricative stimuli are acoustically identical, in principle no purely auditory mechanism can do the job. We must

postulate that phonetic processing is involved at some level in the perception of the place feature.

REFERENCES

- Bailey, P. J., Summerfield, Q., & Dorman, M. On the identification of sine-wave analogues of certain speech sounds. Haskins Laboratories Status Report on Speech Research, 1977, SR-51/52, 1-25.
- Best, C. T., Morrongiello, B., & Robson, R. The perceptual equivalence of two acoustic cues for a speech contrast is specific to phonetic perception. Haskins Laboratories Status Report on Speech Research, 1980, SR-62, this issue.
- Cooper, F. S., & Mattingly, I. G. A computer-controlled PCM system for the investigation of dichotic speech perception. Journal of the Acoustical Society of America, 1969, 46, 115. (Abstract)
- Dorman, M. F. On the identification of sine-wave analogues of CV syllables. Proceedings of the Ninth International Congress of Phonetic Sciences. Copenhagen: University of Copenhagen, 1979.
- Eimas, P. D., Tartter, V. C., Miller, J. L., & Keuthen, N. J. Asymmetric dependencies in processing phonetic features. Perception & Psychophysics, 1978, 23(1), 12-20.
- Harris, K. S. Cues for the discrimination of American English fricatives in spoken syllables. Language and Speech, 1958, 1, 1-7.
- Isenberg, D., & Liberman, A. M. Speech and non-speech percepts from the same sound. Journal of the Acoustical Society of America, 1978, 64, Suppl. No. 1, J20.
- Kling, J. W., & Riggs, L. A. (Eds.) Woodworth and Schlosberg's experimental psychology. New York: Holt, Rinehart & Winston, 1971.
- Ladefoged, P. Preliminaries to linguistic phonetics. Chicago: University of Chicago Press, 1971.
- Liberman, A. M. Duplex perception and integration of cues: Evidence that speech is different from non-speech and similar to language. In Proceedings of the Ninth International Congress of Phonetic Sciences. Copenhagen: University of Copenhagen, 1979.
- Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. The discrimination of speech sounds within and across phoneme boundaries. Journal of Experimental Psychology, 1957, 54, 358-368.
- Miller, J. L. Nonindependence of feature processing in initial consonants. Journal of Speech and Hearing Research, 1977, 20, 510-518.
- Oden, G. C., & Massaro, D. W. Integration of featural information in speech perception. Psychological Review, 1978, 85, 172-191.
- Pollack, I., & Pisoni, D. B. On the comparison between identification and discrimination tests in speech perception. Psychonomic Science, 1971, 24, 299-300.
- Smith, P. T. Feature-testing models and their application to perception and memory for speech. Quarterly Journal of Experimental Psychology, 1973, 25, 511-534.

FOOTNOTES

¹In accordance with common linguistic practice, we will often use the cover term "labial" to refer to both bilabials [p,b,m] and labiodentals [f,v], and the cover term "dental" to refer to both alveolars [t,d,n] and interdental [θ,ð]. Underlying this choice of terminology is the assumption that, at the level of phonological structure, bilabials and labiodentals have the same abstract feature(s) for place of articulation, and likewise alveolars and interdental. This implies, for example, that [s] and [θ], both being voiceless, dental fricatives, would be distinguished at the phonological level by some fourth feature unrelated to place of articulation. In the analysis of Ladefoged (1971), for example, this job is done by the feature [sibilant]. While this type of analysis is usually supported by phonological arguments, it can also be indirectly supported by perceptual evidence: For example, the claim that [s] and [θ] are not distinguished structurally by place, but by a feature like [sibilant], which is related to the character of the frication, is indirectly supported by results like Harris's (1958) observation that [θ] frication combined with [s] transitions gave a [θ] percept, while [s] frication combined with [θ] transitions gave an [s] percept.

²We selected voiced rather than voiceless stops for the comparison because preliminary results showed that truncated fricatives, whether voiced or voiceless, were usually perceived as voiced stops.

³This truncation operation limited our choice of speakers, since for some speakers the friction overlaps with the vocalic part to the extent that a cut after the friction eliminates so much of the transitions that no initial consonant is heard, while a cut at the beginning of the vocalic part leaves so much friction that the initial consonant is unambiguously fricative. The speakers for this experiment were chosen from those where a cut point at the beginning of the vocalic portion (and so presumably including all the voiced transitions) gave a reasonably stop-like percept.

⁴As we had expected, the results for the interdental showed substantial variation from speaker to speaker. For two speakers, TH and PN, [θa]- when perceived as a stop was primarily identified as labial 'ba' (98% 'ba' responses for TH, 91% for PN). For the other two, it was balanced on the labial/dental border (DI: 55% 'ba,' 45% 'da'; PB: 42% 'ba,' 58% 'da'). For all four speakers, however, [θa]- go! significantly more labial responses when it was identified as a stop than when it was identified as a fricative ($p < .01$ by Chi-squared in each case, pooling results for all subjects). Recall that TH and DI are American, while PN and PB are English. This suggests that the inter-speaker variation we see here reflects individual differences rather than regional dialects.

⁵We can test this by looking at the responses of individual subjects to individual tokens. Each subject heard 10 repetitions of each token; suppose we score only those cases where a subject gave at least 2 stop and 2 fricative identifications. The resulting pattern, both for the pooled data and for the individual speakers, looks essentially like the overall pattern: For example, [θa]- with stop manner was identified as labial 69% of the time and as dental 31%; with fricative manner, labial 26% and dental 74%. This difference is

significant ($p < .01$ by Chi-squared for the pooled data; $p < .025$ by Chi-squared for each of the speakers independently).

⁶At the suggestion of David Pisoni this experiment was replicated in its entirety using fewer subjects (8) but more observations per data point (30 instead of 15 for identification, 24 instead of 12 for discrimination) for each subject. The main results of the replication were in accordance with those presented here.

⁷Note that our stimuli are all [Ca] syllables. Preliminary results suggest that the effect we have observed is sensitive to both vowel height and rounding, and to CV vs. VC order.

⁸One discrepancy requires comment: Recall that the identification data from Experiment 3 implied a fricative boundary about 4.35 and a stop boundary about 5.99. This would imply that natural [θ] transitions should correspond to a point on the synthetic continuum around stimulus 5, which does give identification data similar to natural [θa]-. However, the copy-synthesis version of PWJ's natural [θa] is stimulus 7. Given the small absolute difference between stimuli 5 and 7 on the synthetic continuum (e.g., 86 Hz in the starting frequency of F2), the lack of precision in copy synthesis, and the different general character of natural and synthetic stimuli, we do not believe that this discrepancy is evidence against the two-boundary hypothesis.

⁹It is important to notice that this rationalization for our apparent boundary shift cannot be extended to the boundary shifts found by Miller (1977): [p,b,m] are all equally bilabial, and [t,d,n] equally alveolar, so there is no apparent manner/place interaction in articulation to appeal to. It will therefore be important to test whether the effect Miller found works in the same way as ours, in particular whether we can replicate Miller's shifts by manipulating the manner percept while holding the acoustic stimulus constant.

SOME ARTICULATORY CORRELATES OF PERCEPTUAL ISOCHRONY*

Betty Tuller+ and Carol A. Fowler++

Abstract. Given sequences of digits with temporally equidistant acoustic onsets, listeners do not perceive them as isochronous (Morton, Marcus, & Frankish, 1976). For the sequences to be perceptually isochronous, systematic departures from acoustic isochrony must be introduced. These acoustic departures are precisely those that talkers generate when asked to produce an isochronous sequence (Fowler, 1979), suggesting that listeners judge isochrony from acoustic information about articulatory timing. The present experiment was an attempt to test directly whether perceptually isochronous sequences have isochronous articulatory correlates. Electromyographic potentials were recorded from the orbicularis oris muscle when speakers produced sequences of monosyllables "as if speaking in time to a metronome." Sequences were devised so that lip muscle activity was related to the syllable-initial consonant, the stressed vowel, or the stressed vowel and final consonant. Results indicate that isochronous muscle activity accompanies both isochronous and anisochronous acoustic signals produced under instructions to generate isochronous sequences. These results support an interpretation of the perceptual phenomenon reported by Morton et al. (1976) to the effect that listeners judge isochrony of the talker's articulations as they are reflected in the acoustic signal.

INTRODUCTION

Morton, Marcus, and Frankish (1976) report that sequences of digits presented with temporally equidistant acoustic onsets are not perceptually regular. Moreover, when listeners are allowed to adjust the intervals between successive digits, they introduce systematic deviations from acoustic regularity before judging the sequence of alternating digits to be isochronous. These deviations are such that the temporal alignment of a word relative to its neighbors varies with the duration of acoustic energy preceding the acoustically-defined onset of its vowel.¹ Thus, a word pair such as "eight-six," repeated in alternation, tends to be arranged in time so that the acoustic onset-to-onset time is less for "eight-six" than for "six-eight."

*Also in Perception & Psychophysics, 1980, 27, 277-283.

+Also at University of Connecticut, Storrs, Ct., 06268

++Also at Dartmouth College, Hanover, N.H., 03755

Acknowledgment: The research reported here was supported by the following grants to Haskins Laboratories: NINCDS grants NS-13617 and NS-13870, and BRS Grant RR-05596.

Based on these findings, Morton et al. proposed that listeners judge the timing of the word sequences based on reference points within each word. Their data suggest that the longer the acoustic duration of the initial consonant, the longer the interval between the acoustic onset of the word and the location of its reference point. Morton et al. call the reference point the "P-center" of a word, and describe it as the word's "psychological moment of occurrence."

The technique of allowing a listener to adjust the intervals between words only establishes the relative temporal alignments of successive words in a perceptually isochronous sequence; it does not establish the location of the P-center within each word. Moreover, further investigation by Morton et al. did not reveal any acoustic marker of a P-center. Specifically ruled out by their investigations are the stressed vowel's acoustic onset, the word's onset and the vowel's peak intensity.

Two investigations (Allen, 1972; Rapp, Note 1) apparently do locate the P-center in a word, but both fail to discover how it is marked acoustically. Allen's subjects tapped "on the beat" of a designated syllable in a sentence whereas Rapp's subjects repeated nonsense words in time with a regularly occurring pulse. Both studies revealed the same relationship between pulse or tap and the onset of a stressed vowel. The pulse or tap was located near the acoustically-defined onset of the stressed vowel, but preceded it by a duration that correlated positively with the duration of the prevocalic consonant or consonant cluster.

In Rapp's data, it is evident that as the acoustic duration of the prevocalic consonant(s) increased, the beat tended both to precede the acoustic onset of the vowel by longer intervals, and to follow the acoustic onset of the syllable by longer intervals. If stress beats are P-centers, and if P-centers occur at isochronous intervals in perceptually isochronous sequences, then Rapp's data predict the acoustic anisochronies reported by Morton et al. That is, syllables with acoustically long prevocalic consonants should be located temporally closer to their predecessor than syllables with acoustically short prevocalic consonants.

The three investigations by Morton et al., by Allen, and by Rapp, concur in finding that departures from acoustic isochrony are systematic, and that the psychological moment of occurrence, P-center, or stress beat of a word varies in locus with initial consonant duration, but does not correspond to any obvious acoustic marker.

Based on a series of experiments, Fowler (1979) proposed an articulatory rationale for the P-center phenomenon. She found that when talkers are asked to produce an isochronous sequence of stressed monosyllables, they generate an acoustic signal in which the anisochronies are just those that listeners require in order to hear a sequence as isochronous. In itself, this may mean only that talkers and listeners agree in mistaking anisochrony for isochrony. However, an additional study suggests a different interpretation of the findings to the effect that the acoustic anisochronies arise as a by-product of a fairly straightforward strategy for producing isochronous sequences. The study obtained vocal reaction times (as measured from spectrographic displays) for reading visually presented CV syllables. Reading reaction times correlat-

ed negatively with the measured acoustic duration of the initial consonant ($r = -.56$); that is, reaction time was short to syllables with long initial consonants. This is comparable to the finding by Morton et al. that digits with acoustically long initial consonants follow their predecessors more closely in time than digits with short-duration initial consonants. Moreover, differences in reading times between pairs of CVs differing in initial consonant correlated highly with the departures from isochrony of syllables differing in the same consonants observed in the first study ($r = .86$).

It seemed likely that, in the reaction time study, talkers initiated the production of each CV at nearly equivalent intervals following the visual presentation of the CV to be produced. Therefore, the differences in vocal reaction time among the CVs are probably due, for the most part, to the different times after articulatory onset that different consonants have acoustic consequences other than silence. These differences in reaction time correlate highly with deviations from acoustic isochrony in perceptually isochronous utterances. Hence it seems likely that these latter departures, too, are products of isochronous articulatory gestures.

A plausible explanation of the talker's behavior when producing sequences of uniformly stressed isochronous monosyllables, then, is that he or she initiates some aspect of the production of each stressed syllable at isochronous intervals. When initial consonants of successive syllables differ, particularly in manner class, the resulting acoustic signal may be anisochronous. So, for example, in the production of the stop consonant /b/, bilabial closure is maintained to build intraoral pressure for an explosive release. The acoustic correlate of this occlusion typically is silence. In contrast, in the production of /f/, the vocal tract is obstructed by the lip/teeth approximation, but it is not occluded entirely. Therefore, a period of silence need not precede the acoustic onset of an /f/ (or other fricative). Thus, even if lip gestures were initiated at temporally equidistant intervals in a sequence of syllables alternately starting with /b/ and /f/, the acoustic product would be anisochronous.

If psychological isochrony is indeed based on isochrony of articulatory gestures, then we may ask to what articulatory marker the P-center corresponds. Most obviously, it may correspond to a syllable's articulatory onset. However, this account probably fails to explain the data of Rapp and of Allen, cited earlier, which locate the stress beat within the acoustic realization of a consonant. Their findings may signify that articulatory gestures within the syllable, perhaps those associated with the vowel, are the gestures that a talker controls and that a perceiver listens for. Indeed, due to coarticulation among adjacent segments, the gesture for a vowel begins during the production of a prevocalic segment (see, e.g., Gay, 1977; Carney & Moll, 1971). Hence, a proposal that stressed vocalic segments are the timed articulatory gestures underlying the P-center or stress beat effect is not implausible as long as the controlled vocalic gestures are time-locked to those of the consonant.

The present experiment had as its major aim to test directly the proposal that psychologically isochronous sequences have isochronous articulatory correlates, and in doing so, to uncover the articulatory correlate itself. To these ends, we examine muscle activity at the lips during the production of

uniformly stressed isochronous sequences of monosyllables, and compare the relative degrees of isochrony of the muscle activity and of the consequent acoustic signal. In some utterances, the lip activity is associated with the production of the initial consonants in the sequence; in other utterances, it is associated with the production of the stressed vowel or the vowel and final consonant. Utterances have been selected based on earlier results, so that some are expected to be acoustically isochronous, while other tokens are expected to show marked departures from acoustic isochrony. Our concern is to discover whether any part of the production of psychologically isochronous utterances is isochronous--even when the acoustic product departs from isochrony.

METHOD

Subjects

The subjects were four adult females and one adult male. All are native speakers of English; four were naive to the purposes of the experiment.

Stimuli

Two kinds of CVC nonsense-syllable sequences were devised. Alternating sequences included two stressed monosyllables produced in alternation approximately twenty times each. Homogeneous sequences included a single stressed monosyllable repeated about forty times.

Alternating sequences were devised in accordance with two criteria. First, based on earlier studies, the spoken sequences were expected to be acoustically anisochronous; second, both component syllables of a sequence involve the orbicularis oris muscle for articulation either of the initial consonant or of the stressed vowel. (The orbicularis oris acts to close or protrude the lips, and is associated with the production of labial consonants and lip rounded vowels [e.g., Bell-Berti & Harris, 1979].) Three alternating sequences were selected:

1. /bak fak bak fak.../ In this sequence, the orbicularis oris is associated with the gestures for /b/ and /f/. Because /f/ is a fricative and /b/ is a stop, the acoustic departures from isochrony should be such that the interval from /bak/ to /fak/ is less than that between /fak/ and /bak/.
2. /duk suk duk suk.../ In this utterance, orbicularis oris activity is associated with lip rounding for the vowel /u/. Acoustically, the /duk/-/suk/ interval should be shorter than the /suk/-/duk/ interval.
3. /dup sup dup sup.../ In this utterance, orbicularis oris activity is associated with the vowel and the final consonant. Acoustically, the /dup/-/sup/ interval should be shorter than the interval between /sup/ and /dup/.

Homogeneous utterances were selected to match the alternating sequences. They were /bak bak.../, /duk duk.../, /suk suk.../, /dup dup.../, /sup sup.../. (In initial acoustic measurements of /bak fak.../, the low amplitude

of /f/-frication made it very difficult to obtain accurate measurements of acoustic onset; hence the homogeneous sequence /fak fak.../ was not used in the experiment. However, /fak/ was used in the alternating utterances to provide a labial fricative of long acoustic duration in contrast to the stop, /b/.)

Measures

Electromyographic (EMG) activity was recorded from the orbicularis oris muscle using paint-on surface electrodes (Allen & Lubker, 1972) spaced at approximately half a centimeter on the vermilion border of the lips. The EMG potentials and acoustic recordings of the subjects' utterances were recorded on separate channels of a multichannel tape.

Procedure

Subjects were asked to produce six sequences of nonsense syllables stressing every syllable equally and producing the syllables at isochronous intervals--that is, as if speaking in time to a metronome. Each subject produced all three alternating sequences and three matching homogeneous sequences (/bak bak.../, /duk duk.../ or /suk suk.../, and /dup dup.../ or /sup sup.../). The order of the syllables in alternating sequences was varied across subjects, some subjects beginning the sequences with the stop- and some with the fricative-initial syllable. Although subjects were not instructed to relax the rounding at the lips between syllables in sequences containing the vowel /u/, activity in orbicularis oris did return to baseline between syllables.

Data Analysis

Each homogeneous and alternating sequence was considered to contain two interval-types, odd and even. Intervals of both types extended from the acoustic onset of the final consonant release of one syllable (the "left margin" of the interval) and the acoustic or EMG onset of the immediately following syllable (the "right margin"; see Figure 1). Among the alternating sequences, odd and even intervals differ in respect to their left and right margins. For example, in /bak fak bak fak.../, odd intervals begin at the release of /k/ in /bak/ and end at the acoustic or electromyographic onset of /f/ in /fak/. Even intervals begin at the release of /k/ in /fak/ and end at the onset of /b/ in /bak/. Homogeneous utterances likewise were partitioned into odd and even intervals, but, now, the intervals were identical in respect to their left and right margins.

In the case of two subjects, sequences ending in /p/ were not released reliably. For these utterances, the offset of periodicity in the preceding vowel was used as the left margin of odd and even intervals.

To avoid contamination of the isochrony measurements by utterance-initial and -final lengthening effects (Klatt, 1976; Lindblom & Rapp, Note 2; Oller, 1973), the first two syllables and the last two syllables of all sequences were excluded from the analyses. The number of tokens of each interval-type, odd and even, ranged from 15 to 20 for each subject.

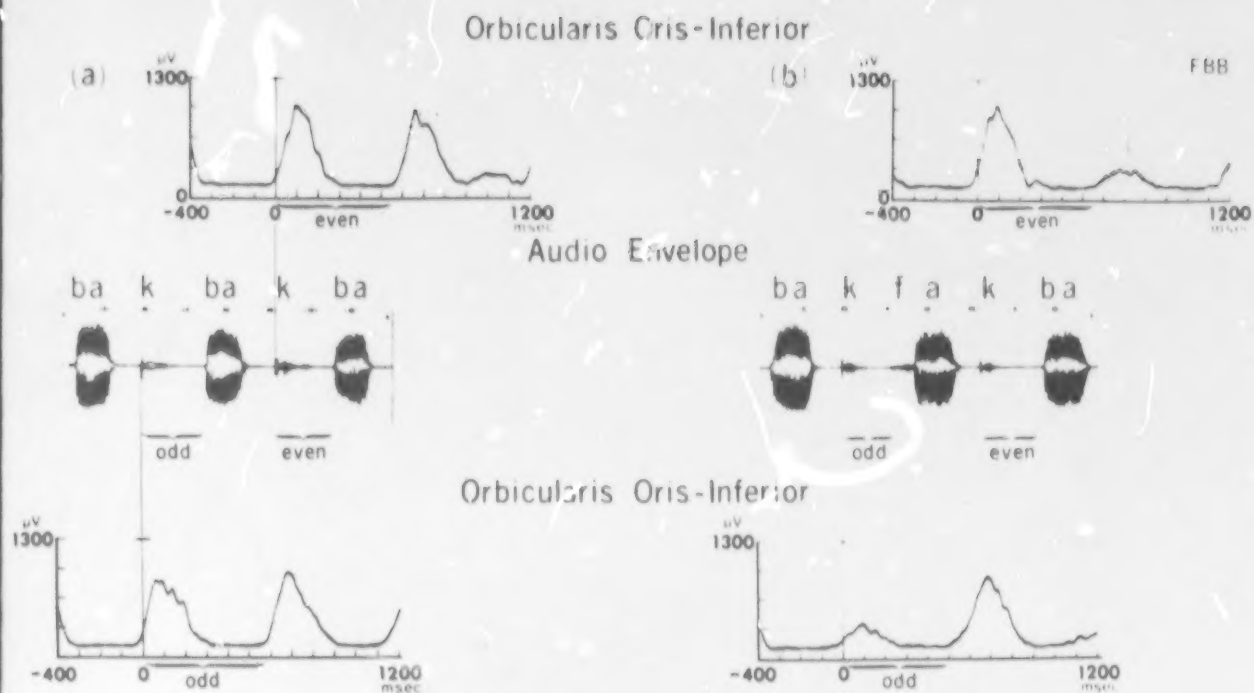


Figure 1. "Odd" and "even" intervals in acoustic and EMG records for both a) homogeneous and b) alternating sequences.

Acoustic measurements were made from sound spectrograms. The EMG data were rectified, software integrated with a time constant of 35 msec and averaged using the Haskins Laboratories EMG system described by Kewley-Port (1973, 1974). The left margin of each interval, as defined above, was used as the point around which tokens of each interval-type were averaged.

The averaging program provides a listing of the mean amplitude of the EMG signal in microvolts during successive 5 msec intervals throughout the averaged interval. Using this listing, the onset of EMG activity in the orbicularis oris muscle following consonant release was defined as the point in time at which 5% of the amplitude range between baseline and peak activity on the averaged signal had been traversed. Typically, no reversals of the EMG signal occurred beyond that point.

We emphasize that the EMG onsets were taken from signals averaged across all of the odd or even intervals within an utterance as produced by a single talker. That is, they were not taken from individual odd or even intervals. Whereas onsets are easily identified from averaged signals, they cannot be identified reliably from raw signals (Baer, Bell-Berti, & Tuller, 1979). This procedure of identifying onsets from an averaged signal has the disadvantage that in contrast to our acoustic measures, variances cannot be obtained on individual utterances spoken by individual talkers. Instead, they must be obtained across utterances, or, as we have done, across talkers.

RESULTS

For both acoustic and EMG measurements, difference scores were calculated between the means of the two interval types. (The difference scores for /suk suk.../ and for /duk duk.../ did not differ from each other and therefore were combined across subjects, as were those for /sup sup.../ and /dup dup.../.) Let us first consider the acoustic analyses as indicated in Table 1.

Table 1

Mean difference scores (and sd in parentheses) between odd and even intervals of each sequence. Acoustic and EMG measurements are indicated in separate columns.

<u>Sequence</u>	<u>Acoustic</u>		<u>EMG</u>	
	<u>Mean</u>	<u>sd</u>	<u>Mean</u>	<u>sd</u>
/bak bak.../	2.00	(3.79)	13.17	(7.71)
/bak fak.../	-95.00*	(15.27)	-3.00	(11.08)
/duk duk.../	-12.67	(9.66)	1.33	(15.06)
/duk suk.../	-84.17*	(17.20)	-27.00	(24.29)
/dup dup.../	-5.67	(5.01)	-5.50	(11.58)
/dup sup.../	-99.17*	(14.81)	1.17	(26.73)

*p < .01

For all homogeneous utterances, the difference scores are not significantly different from zero. In contrast, among the alternating sequences, the acoustic difference scores depart from zero in all instances. As predicted, the differences are such that intervals from release of /k/ or /p/ to onset of the fricative are shorter than those between /k/ or /p/ release and stop onset. These results hold for each subject considered individually as well as for the pooled data of Table 1. These two results confirm the earlier work on isochrony in utterances of this type: Homogeneous utterances are acoustically isochronous, but alternating utterances depart from isochrony in predictable ways (Fowler, 1979).

The results of the analyses on the EMG data are also given in Table 1. In these data, no difference scores depart significantly from zero either among the homogeneous or among the alternating sequences.

Figure 2 provides an example of a pair of homogeneous and alternating sequences that display the difference in outcome for the acoustic and EMG measures of isochrony. The figure shows clearly that large acoustic departures from isochrony are obtained in the utterance /dup sup.../, but not in /sup sup.../. The EMG measure of lip rounding and of lip activity for the final /p/ are isochronous in both utterances.

A three-way repeated-measures analysis of variance (measure [EMG vs. acoustic] by sequence type [homogeneous vs. alternating] by utterance [bak vs. duk vs. dup]) was computed on the difference scores. The effects of measure, $F(1,4) = 13.9$, $p < .025$, of sequence type, $F(1,4) = 15.75$, $p < .025$, and of their interaction, $F(1,4) = 9.29$, $p < .05$, were significant. No other effect reached statistical significance.

As Table 1 makes clear, the interaction is significant because acoustic difference scores are substantially larger than EMG difference scores for the alternating utterances, but are not larger for homogeneous sequences.

The absence of an effect of utterance-type indicates that for the utterances of this experiment, articulatory activity for the initial consonant is no more nor less isochronous than activity related to the production of the stressed vowel or stressed vowel and final consonant.

DISCUSSION

These results support an interpretation of the P-center phenomenon that invokes the production strategy of the talker as it is reflected in the acoustic signal. In particular, they suggest that, when asked to produce isochronous monosyllabic utterances, talkers comply by producing isochronous articulatory gestures. Due at least in part to differences in the times after articulatory onset that different consonants have acoustic consequences other than silence, the acoustic product of isochronous gestures may depart substantially from isochrony.

The present experiment did not achieve its aim to identify an articulatory P-center. That is, it did not distinguish the production of the initial consonant, the vowel or even the final consonant in degree of articulatory isochrony. The articulation of all of these segments was isochronous in the

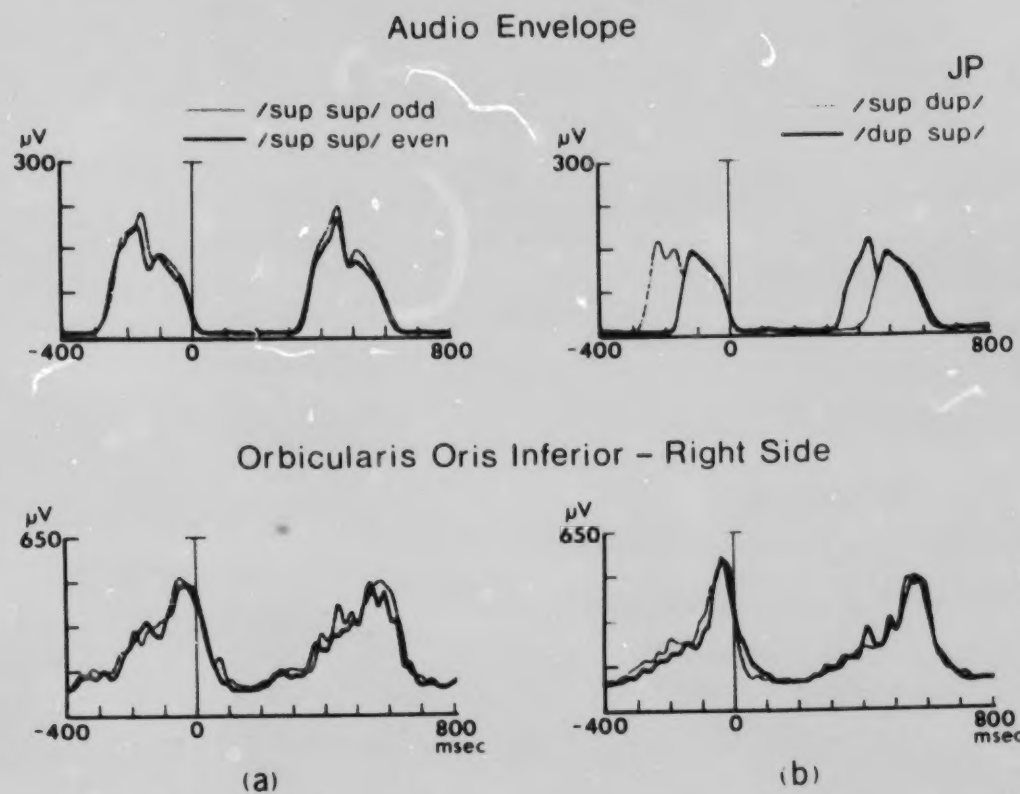


Figure 2. Audio envelope and EMG records for the /sup sup.../ and /dup sup.../ sequence pair for subject JP.

simple utterances of the present experiment. We suspect, however, based on some hints in the literature and on a recent study (Fowler & Tassinary, Note 3) that further investigation will distinguish them. In particular, as noted, Rapp (Note 1) and Allen (1972) both place the regulated "beat" within the acoustic realization of a consonant. Unless there is some consistent tendency to postpone the beat or tap involved in these techniques, the results they provide would tend to rule out the initial consonant of a syllable as the timed event detected by listeners in isochronous productions.

In addition, a recent investigation suggests that the initial segments of alternating utterances are isochronous only if they participate in consonant clusters of the same size (Fowler & Tassinary, Note 3). In that study, acoustic measurements were made of alternating sequences in which the initial consonantal context included clusters and single consonants (e.g., /sad strad sad strad.../ or /tad trad tad trad.../). These sequences showed systematic departures from acoustic isochrony. Specifically, syllables starting with a cluster were abutted more closely to their predecessors than those starting with a single consonant. Since the syllables beginning with clusters in this study had the same initial consonants as those starting with single consonants, it is unlikely that the articulatory onsets of these syllables were temporally equidistant. In contrast to the initial consonants, the vowels apparently remained isochronous independent of cluster size.

These preliminary findings suggest that further investigation of syllable internal gestures, perhaps those associated with the stressed vowel, are most likely to uncover the regulated events in isochronous productions.

In conjunction with the present findings, the perceptual data of Morton et al. (1976) and of Fowler (1979) suggest that listeners' judgments of isochrony are, in some sense, articulation-referential. We briefly consider alternative explanations for these kinds of judgments.

One view of speech perception that explains its articulation-referential character is the motor theory (e.g., Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). Indeed, the motor theory was devised to explain data of precisely the present sort in which descriptions of the articulatory generation of speech segments seem to correspond more closely to a listener's identifications of the segments than do descriptions of their acoustic realizations. These findings include the perceptual equivalence of the rising and falling F2 transitions in synthetic versions respectively of /di/ and /du/ (Liberman, Delattre, Cooper, & Gerstman, 1954), the perceptual "nonequivalence" of acoustically identical stop bursts in the synthetic (Liberman, Delattre, & Cooper, 1952) and real (Schatz, 1954) syllables, /pi/, /ka/ and /pu/, various trading relations that obtain among acoustically quite different "cues" for the same phonetic segment (Fitch, Halwes, Erickson, & Liberman, 1979) and the role of silence in the appropriate contexts as information for stop and affricate manners of articulation (see Liberman & Pisoni, 1977, for a review). The motor theory explains these findings by proposing that listeners access their own production systems in order to identify the sounds of speech.

The motor theory is perhaps unique among views of speech perception in accounting for findings of the foregoing sort. However, because it invokes access to mechanisms for producing the stimulation being perceived, the motor

theory implies that the processing underlying speech perception is unlike that underlying auditory nonspeech perception or perception in any other sensory modality. In respect to this proposal, the theory is somewhat unparsimonious and unattractive.

In fact, the data that the motor theory was devised to explain provide no evidence at all relating to how articulation-referential percepts are achieved by the perceptual system. That is, they do not support a theory that invokes access to the production system over one that does not. Instead, they reveal only what in an acoustic speech signal is perceptually significant--namely, acoustic patterns that reflect articulation, rather than any other acoustic "patterns" that might appear salient to an investigator.

Moreover, in contrast to the view of motor theorists that the articulation-referential character of speech perception makes it unique among perceptual modes, it is possible instead that articulation-referential perception is symptomatic of the close alliance between the perception of speech and other kinds of perception. That is, insofar as it indicates a sensitivity on the part of perceivers to information in proximal stimulation (the acoustic signal in speech or in nonspeech audition, reflected light in vision, patterns of deformation on the skin in touch) that corresponds to or reflects properties of the distal source of the stimulation (the moving vocal tract in speech, a visible or tangible object in vision and touch respectively), the articulation-referential character of speech perception may be analogous to the object-referential character of vision, touch and, perhaps, nonspeech audition. That is, independent of modality, perception seems generally to involve the pick-up of patterns in proximal stimulation that correspond to or reflect aspects (audible, visible, tangible) of distal sources of stimulation.

Explanations as to how this is achieved in perception differ in fundamental ways (e.g., Gibson, 1966; Lindsay & Norman, 1977), but no explanation other than the motor theory invokes access by the perceiver to the means of producing the proximal stimulation. We presume, therefore, that other explanations for articulation-reference in speech perception are possible.

Given that the present data do not provide evidence relating to the means by which perceivers achieve their articulation-referential percepts, we will not consider this issue further. However, the data do provide some indications concerning what in an acoustic speech signal provides information about isochrony to a perceiver and what does not. We consider that issue briefly.

We presume, based on the present findings, that the controlled intervals in isochronous productions are articulatory, and hypothesize that the edges of these intervals are the P-centers of successive syllables, as first described by Morton et al. (1976). If the P-center corresponds to an articulatory event to which a perceiver is sensitive, then some acoustic evidence of the P-center must mark its location for him. For example, if the P-center corresponds to the articulatory onset of a stressed vowel, then further investigation should show that it corresponds also to some of the detectable acoustic correlates of the onset in the speech signal. One correlate of the articulatory onset of a vowel is the onset of its influence on the acoustic realization of the consonant. Similarly, if the P-center is found to correspond to the articulatory onset of the syllable-initial or -final consonant, or to the closest

approximation of a vowel to its "target" shape of the vocal tract (cf. MacNeillage, 1970), then whatever the acoustic correlates of those articulatory events might be, they should be found to correspond to the perceived locus of the P-center.

We should point out that none of these proposed articulatory markers has been shown to correspond to the P-center, and, indeed, the literature offers apparent counter-evidence to some of them. For example, as we have pointed out, the articulatory onset of the syllable-initial consonant occurs consistently earlier than the syllable's P-center. For its part, the onset of the syllable-final consonant occurs consistently later than the syllable's P-center. One likely candidate may be the vowel's attainment of its target shape of the vocal tract; although the articulatory onset of the vowel occurs earlier in the consonant than the P-center (as inferred from the data of Carney & Moll, 1971; Gay, 1977; see also Bell-Berti & Harris, 1979; Bell-Berti, Harris, Tuller, Baer, & D'Antonio, 1978), any detectable acoustic correlates of the vowel's onset may occur somewhat later.

We point out in conclusion that we have demonstrated articulatory isochrony, but only under restricted conditions. First, we have varied only initial consonants in the alternating sequences, and have considered only the stop and fricative manner classes of consonant. To verify that some aspect of articulation remains isochronous in any psychologically isochronous sequence of stressed monosyllables, other variations in initial consonant (including clusters, for example), in the stressed vowel and in the final consonant require examination. Second, the acoustic departures from isochrony that we have considered are only those due to P-center effects (and, thus, only segmental effects). Larger departures from acoustic isochrony occur in perceptually stress-timed utterances that contain both stressed and unstressed syllables (e.g., Lehiste, 1973; see also Lehiste, 1977; Donovan & Darwin, Note 4). The conclusions of the present study cannot be generalized to these departures from isochrony without investigation of utterances that include unstressed syllables. Despite these limitations in scope, however, we have demonstrated the occurrence of invariant articulatory timing that covaries with the judgments of listeners under conditions in which intervals based on articulation-free acoustic measures show significant variability.

REFERENCE NOTES

1. Rapp, K. A study of syllable timing. Papers from the Institute of Linguistics, University of Stockholm, 1971, 8, 14-19.
2. Lindblom, B., & Rapp, K. Some temporal regularities of spoken Swedish. Papers from the Institute of Linguistics, University of Stockholm, 1973, 21, 1-59.
3. Fowler, C. A., & Tassinary, L. Stress timing in speech production and perception. Manuscript in preparation.
4. Donovan, A., & Darwin, C. The perceived rhythm of speech. Paper presented at the IXth International Congress of Phonetic Sciences, Copenhagen, August, 1979.

REFERENCES

- Allen, G. The location of rhythmic stress beats in English, I. Language and Speech, 1972, 15, 72-100.
- Allen, G. D., & Lubker, J. F. New paint-on electrodes for surface electromyography. Journal of the Acoustical Society of America, 1972, 52, 124. (Abstract)
- Baer, T., Bell-Berti, F., & Tuller, B. On determining EMG onset time. In J. J. Wolf & D. H. Klatt (Eds.), Speech Communication Preprints, 97th Meeting of the Acoustical Society of America, Cambridge, Mass., 1979. New York: Acoustical Society of America, 1979.
- Bell-Berti, F., & Harris, K. S. Anticipatory coarticulation: Some implications from a study of lip rounding. Journal of the Acoustical Society of America, 1979, 65, 1268-1270.
- Bell-Berti, F., Harris, K. S., Tuller, B., Baer, T., & D'Antonio, L. Some experiments on anticipatory coarticulation. Journal of the Acoustical Society of America, 1978, 64, S92. (Abstract).
- Carney, P., & Moll, K. A cinefluorographic investigation of fricative consonant-vowel coarticulation. Phonetica, 1971, 23, 193-202.
- Fitch, H., Halwes, T. G., Erickson, D., & Liberman, A. Perceptual equivalence of two acoustic cues for stop-consonant manner. Haskins Laboratories Status Report on Speech Research, 1979, SR-57, 183-200.
- Fowler, C. "Perceptual centers" in speech production and perception. Perception & Psychophysics, 1979, 25, 375-388.
- Gay, T. Articulatory movements in VCV sequences. Journal of the Acoustical Society of America, 1977, 62, 183-193.
- Gibson, J. J. The senses considered as perceptual systems. Boston, Mass: Houghton-Mifflin, 1966.
- Kewley-Port, D. Computer processing of EMG signals at Haskins Laboratories. Haskins Laboratories Status Report on Speech Research, 1973, SR-33, 173-183.
- Kewley-Port, D. An experimental evaluation of the EMG data processing system: Time constant choice for digital integration. Haskins Laboratories Status Report on Speech Research, 1974, SR-37/38, 65-72.
- Klatt, D. The linguistic uses of segment duration in English. Journal of the Acoustical Society of America, 1976, 59, 1208-1221.
- Lehiste, I. Rhythmic units and syntactic units in production and perception. Journal of the Acoustical Society of America, 1973, 54, 1228-1234.
- Lehiste, I. Isochrony revisited. Journal of Phonetics, 1977, 5, 253-263.
- Liberman, A., Cooper, F., Shankweiler, D., & Studdert-Kennedy, M. Perception of the speech code. Psychological Review, 1967, 74, 431-461.
- Liberman, A., Delattre, P., & Cooper, F. The role of selected stimulus variables in the perception of the unvoiced stop consonants. American Journal of Psychology, 1952, 65, 497-516.
- Liberman, A., Delattre, P., Cooper, F., & Gerstman, L. The role of consonant-vowel transitions in the perception of the stop and nasal consonants. Psychological Monographs, 1954, 68 (Whole No. 379).
- Liberman, A., & Pisoni, D. Evidence for a special speech-perceiving subsystem in the human. In T. Bullock (Ed.), Recognition of complex acoustic signals. Berlin: Dahlem Konferenzen, 1977.
- Lindsay, P. H., & Norman, D. Human information processing (2nd ed.). New York: Academic Press, 1977.

- MacNeilage, P. Motor control of serial ordering of speech. Psychological Review, 1970, 77, 182-196.
- Morton, J., Marcus, S., & Frankish, C. Perceptual centers. Psychological Review, 1976, 83, 405-408.
- Oller, D. The effects of position in utterance on speech segment duration. Journal of the Acoustical Society of America, 1973, 54, 1235-1246.
- Schatz, C. The role of context in the perception of stops. Language, 1954, 30, 47-56.

FOOTNOTES

In this paper, we will use expressions such as "acoustic onset of the vowel" or "acoustic onset of the consonant." These phrases are loose ways of characterizing acoustic, but not necessarily linguistic or articulatory boundaries in an optical display of speech. For example, "acoustic onset of the vowel" typically means the onset of periodic energy for the vowel. This is a conventional, but, in fact, an incorrect way to mark a vowel's onset. Due to coarticulation, the articulatory onset of a vowel may occur in a preceding consonantal segment, and its acoustic influence on the signal will begin there. The so-called "vowel onset," defined as the onset of periodic energy, is the time at which the vowel predominates in its influence on the character of the acoustic signal--an event that may be of no psychological significance whatever. We begin by using these phrases, and by taking our measures with reference to the boundaries they indicate because they are the measures that have been used to characterize the P-center or stress-beat. Eventually, however, we will want to revise our descriptions and our measurements to make them compatible with the acoustic markers that are psychologically significant.

EFFECTS OF UTTERANCE CONTINUITY ON PHONETIC JUDGMENTS*

David R. Dechovitz+, Brad Rakerd+, and Robert R. Verbrugge+

Abstract. The several acoustic cues to a phonetic segment are generally distributed over broad stretches of the speech signal. The present research suggests that discontinuities in the production of speech can constrain the stretch of signal over which these features are integrated in phonetic perception. In one experiment, a period of silence was shown to distinguish affricate from fricative after precursors that were ongoing utterances but not after precursors that were completed utterances. This was so both when precursors and target were produced by the same talker and when they were produced by different talkers. In a second experiment, this effect was shown to hold over a wide range of silent intervals and, in the final experiment, it was shown that the effect could not be accounted for by listeners' perceptions of the number of talkers who had produced the utterances. It is argued that listeners are sensitive to information specifying continuity of articulation, that this sensitivity can influence phonetic perception, and further, that the information for continuity is partially independent of the information for vocal tract source.

The several acoustic cues contributing to a single phonetic percept are generally distributed over broad stretches of the speech signal. For example, an intervocalic stop consonant is typically marked acoustically by rising and falling transitions of the several formants, by a period of relative silence, and then by a second set of formant transitions. Since the features relevant for the perception of a particular phonetic segment are spread out in this way, they must be integrated over time, and a number of studies have been taken to show that the process of perceptual integration is carried out with reference to speech articulation (Bastian, Eimas, & Liberman, 1961; Repp, Liberman, Eccardt, & Pesetsky, 1978; Dorman, Raphael, & Liberman, 1979). That is, the listener is said to integrate just those acoustic features that derive from a single phonetic articulation.

In one study compatible with this view, it was shown that insertion of an appropriate duration of silence following say, in a natural production of

*Portions of this paper were presented at the 96th meeting of the Acoustical Society of America, Honolulu, Hawaii, 27 November to 1 December 1979.

+Also University of Connecticut, Storrs.

Acknowledgment. This research was supported by NICHD grant HD-01994 and BRS grant RR-05596 to Haskins Laboratories.

[HASKINS LABORATORIES: Status Report on Speech Research SR-62 (1980)]

Please say shop, yields a pattern identified as Please say chop, and complementarily, that removal of the silent interval from a natural production of Please say chop yields a pattern heard as Please say shop (Dorman et al., 1979). The authors' account of these results is that in intervocalic position silence informs the listener that vocal tract closure was effected by a talker (since no sound is emitted for the period of closure), and this closure distinguishes the articulation of the affricate /ʃ/ from that of the fricative /ʃ/.

By this account, silence serves as an effective cue because it carries information about phonetic articulation. Note, however, that silence may result not only from articulatory closure, but also from a discontinuity in articulation, such as occurs when a talker pauses to breathe. If a listener interprets silence in light of its articulatory implications, it follows that silence should only convey phonetic information when it occurs in a stretch of speech which is perceived to have been produced continuously. Evidence for this has been provided by a demonstration that silence had no phonetic significance when two talkers produced the utterance Please say chop (one producing Please say and the other shop) (Dorman et al., 1979). In this case, a discontinuity arising from a change of talkers undermined the phonetic significance of the silence cue.

A break in the fluent articulation of a single vocal tract source might be expected to have similar perceptual consequences. Such a break might arise if a talker produced a target word and precursor separately. There is evidence, for example, that speech production is organized in terms of units that respect both syntactic boundaries in the language (Boomer & Laver, 1968; Cooper, Lapointe, & Paccia, 1977; Cooper, Egido, & Paccia, 1978) and physiological constraints on respiration (Lieberman, 1967). If adjacent stretches of speech were produced in separate articulatory units, then one would not expect to find a phonetically significant gesture that spans them; consequently, a silent interval between two such units, even if it is of appropriate duration, might not be interpreted as a stop closure. In order to test this prediction, we have assessed the role of silence in distinguishing fricative from affricate following precursors that were spoken as completed utterances and following precursors that were excised from ongoing utterances. To permit a comparison of these results with those observed at junctures where the vocal tract source is changed, we have included a two-talker condition (cf. Dorman et al., 1979).

EXPERIMENT 1

Method

In a sound-attenuated room a male talker produced the statements Let's go and Let's go shop, and the question Let's go? in time with a metronome. He was instructed to produce each utterance in a single breath. A female talker, having listened to the male's productions, mirrored his stress patterns while producing the two statements at the same metronomic rate and with the same respiratory constraint. Five utterances of the word pair let's go were thus generated. Three of these utterances ~~were~~^{were} spoken in isolation. These included the male talker's Let's go statement, his Let's go? question, and

the female talker's Let's go statement. These will be referred to as completed utterances, to reflect the fact that their articulation had been completed prior to the production of any succeeding words. The two remaining productions of let's go contrasted with the completed utterances in that they had been articulated continuously with a following word (and were later edited out for use in this study). These will be referred to as truncated utterances, to reflect their interruption in the midst of an ongoing articulation. These included the Let's go/ excised from the male talker's recording of Let's go shop, and Let's go/ excised from the female talker's recording of Let's go shop.

Whatever the precise units of articulation may be, it would be expected that they are closed at the end of the completed utterances, since each of these utterances was produced in a single breath and each was intended to convey a complete syntactic constituent. By contrast, articulation should not have been closed at the end of the truncated utterances, since each reflected the interruption of articulation in the midst of a breath and prior to a major constituent boundary.

Oscillographic comparison of these utterances demonstrates that they were similar in overall duration, and in the timing of the onset and duration of their constituent words (items a, b, d, e, & f in Figure 1). The overall amplitude envelopes were similar as well; however, the amplitude level at the end of the male talker's completed-utterances was appreciably lower than that at the end of his truncated-utterance. To control for this amplitude difference, an additional utterance was created in which the amplitude envelope of his completed-utterance statement was modified to approximate that of the truncated production. This modified completed-utterance statement is displayed as item c in Figure 1. These six utterances made up the set of precursors.

The male talker's shop production was excised from his recording of Let's go shop immediately following the final pitch pulse of go, and appended to each of the six precursors following 75 msec of silence. (Informal listening had indicated that the 75 msec interval was sufficient for the shop token to be heard as chop.) The shop token was also converted to a token heard as chop in isolation. This was accomplished by deletion of its first 20 msec and by deletion of an additional 30 msec of more medial portions of the frication, thereby creating an abrupt onset and shortened overall duration which are cues for the affricate (cf. Gerstman, 1957). The chop token was then appended to the six precursors without intervening silence, the absence of silence having previously been shown to be sufficient for chop to be heard as shop in the context of continuous speech from a single talker (Dorman et al., 1979).

A total of 12 stimuli were produced, each a version of the three-word string let's go shop. Eight of the 12 were one-talker stimuli. Four of these eight paired the male talker's shop token with his truncated-utterance precursor and with his three completed-utterance precursors (the statement, the modified statement, and the question), and the four remaining one-talker stimuli paired the chop token with these same precursors. The four two-talker stimuli paired the male talker's shop and chop tokens with the female's truncated-utterance and completed-utterance precursors.

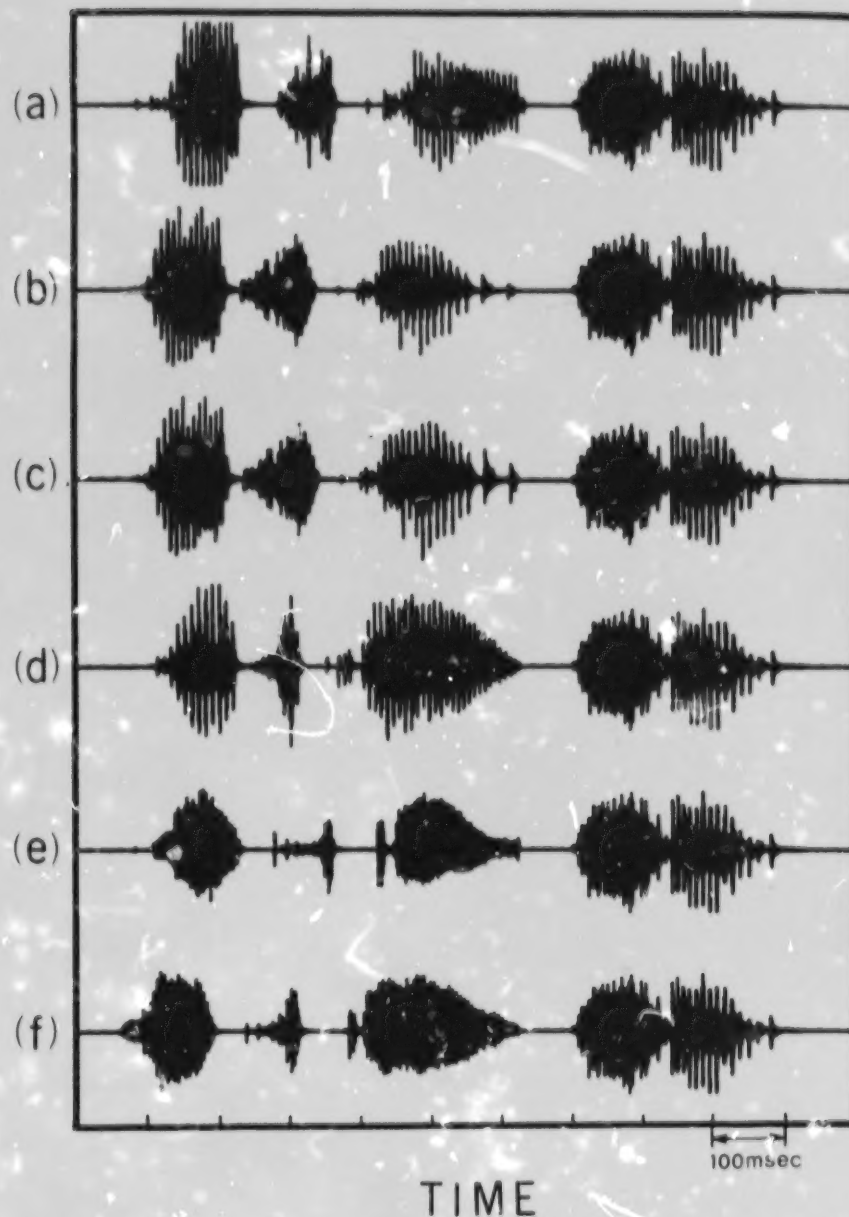


Figure 1. Oscillograms of the shop-token stimuli aligned at the point of silent interval onset. The precursors are: (a) the one-talker truncated-utterance statement, (b) the one-talker completed-utterance statement, (c) the one-talker completed-utterance modified statement, (d) the one-talker completed-utterance question, (e) the two-talker truncated-utterance statement, (f) the two-talker completed-utterance statement.

Ten repetitions of each of these 12 stimuli were randomized and presented over headphones at a comfortable listening level. Listeners were tested in groups of five in a sound-attenuated room, with instructions to identify the final word in each three-word stimulus as either shop or chop and to indicate their choices by writing s or c on an answer sheet.

It was evident that several of the listeners were insensitive to the silence cue in all of the experimental conditions. These listeners failed to demonstrate even the single-talker silence effects reported by Dorman et al. (1979). Criteria were therefore established to screen listeners before data analysis. Listeners were included in all analyses of the shop stimuli if they identified the shop-token target as chop on at least 50% of the trials in which it was paired (following 75 msec of silence) with the one-talker truncated-utterance precursor. Likewise, listeners were included in all analyses of the chop stimuli if they identified the chop-token target as shop on at least 50% of the trials in which it was paired (following 0 msec of silence) with the one-talker truncated-utterance precursor. Of 29 listeners tested, 24 met the criterion for the shop-token stimuli and 12 met the criterion for the chop-token stimuli.

Results

Recall that the truncated-utterance precursors were excised from ongoing utterances, while the completed-utterance precursors were spoken in isolation, and that these precursors were then paired with the shop-token target (following 75 msec of silence) or with the chop-token target (following 0 msec of silence) in a perceptual test. The results of this test are summarized in Tables 1 and 2, with Table 1 displaying the percentage of chop responses given to the shop-token stimuli and Table 2 displaying the percentage of shop responses given to the chop-token stimuli. These measures are indices of the degree to which phonetic perception was influenced by the presentation of a target in the context of a particular precursor and silent interval. In Table 1, for example, the measure reflects the degree to which an acoustic pattern heard as shop in isolation was identified as chop in context.

It can be seen that the silence manipulation reliably influenced phonetic judgments following truncated-utterance precursors and that it had little influence following completed-utterance precursors. This difference was highly significant, both for the shop-token stimuli, $F(1,23)=348.33$, $p<.001$, and for the chop-token stimuli, $F(1,11)=52.17$, $p<.001$. In addition, for the chop-token stimuli there was a significant main effect of the number of talkers (one or two) producing an utterance, $F(1,11)=11.15$, $p<.01$, and an interaction of precursor type (truncated or completed) and number of talkers, $F(1,11)=10.89$, $p<.01$.

There was no significant difference ($p<.05$) among the three one-talker completed-utterance precursors for either the shop-token stimuli, $F(2,46)=2.65$, or the chop-token stimuli, $F(2,22)=0.04$. Therefore, in Figures 2 and 3 only the results for the completed utterance statement (Let's go) are displayed for comparison with the stimuli from the other one- and two-talker conditions. Again, it can be seen that the precursor manipulation had a striking influence on responses to both the shop and chop tokens, while the

Table 1

Mean Percentage of Chop Responses to the Shop-token Target
When Paired with Precursors and a 75 msec Silent Interval

<u>Precursors</u>	<u>Number of Talkers</u>	
	<u>One</u>	<u>Two</u>
Truncated-Utterance Statement	85.8	85.0
Completed-Utterance*		
Statement	15.8	17.1
Modified Statement	12.5	
Question	10.4	

* A mean score for the three completed-utterance one-talker conditions was computed for each listener ($n=24$) and these scores were used in the analysis of variance comparing the effects of precursor type and number of talkers.

Table 2

Mean Percentage of Shop Responses to the Chop-token Target
When Paired with Precursors and a 0 msec Silent Interval

<u>Precursors</u>	<u>Number of Talkers</u>	
	<u>One</u>	<u>Two</u>
Truncated-Utterance Statement	89.2	59.2
Completed-Utterance*		
Statement	10.8	10.0
Modified Statement	10.8	
Question	11.7	

* A mean score for the three completed-utterance one-talker conditions was computed for each listener ($n=12$) and these scores were used in the analysis of variance comparing the effects of precursor type and number of talkers.

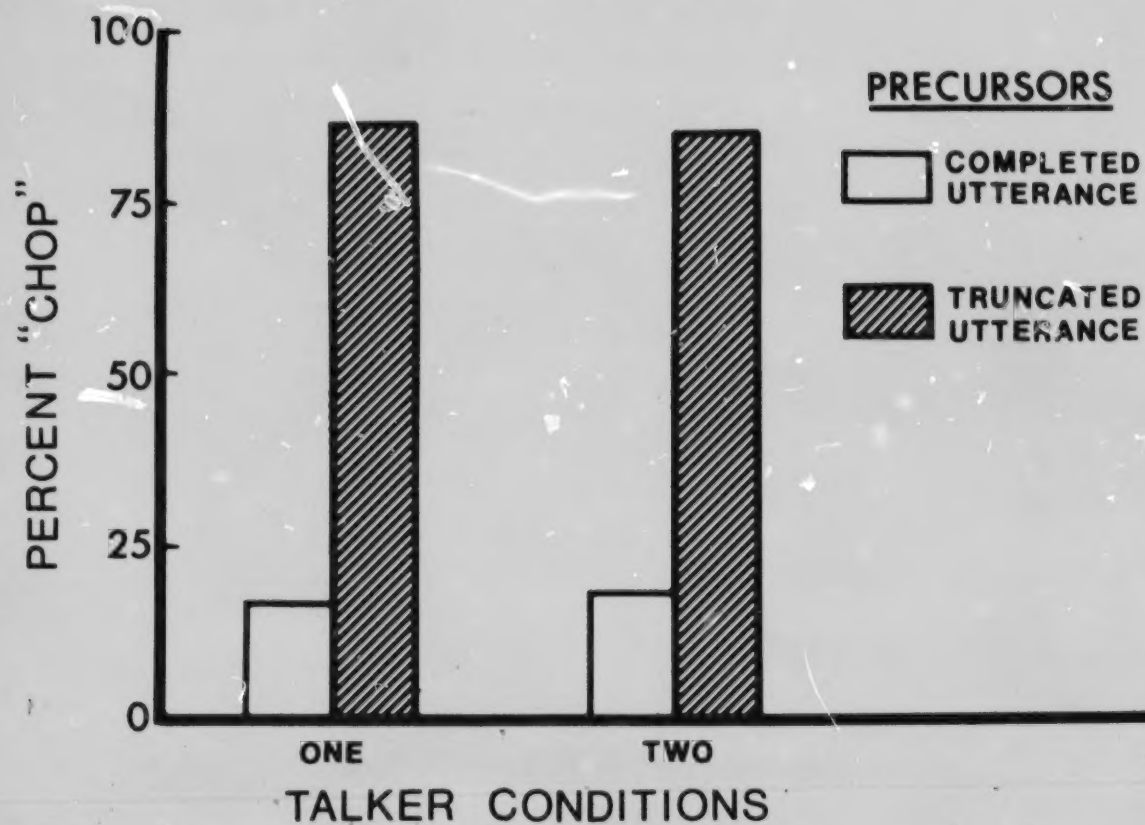


Figure 2. Percentage of chop responses for stimuli pairing different precursors with 75 msec of silence and the shop token.

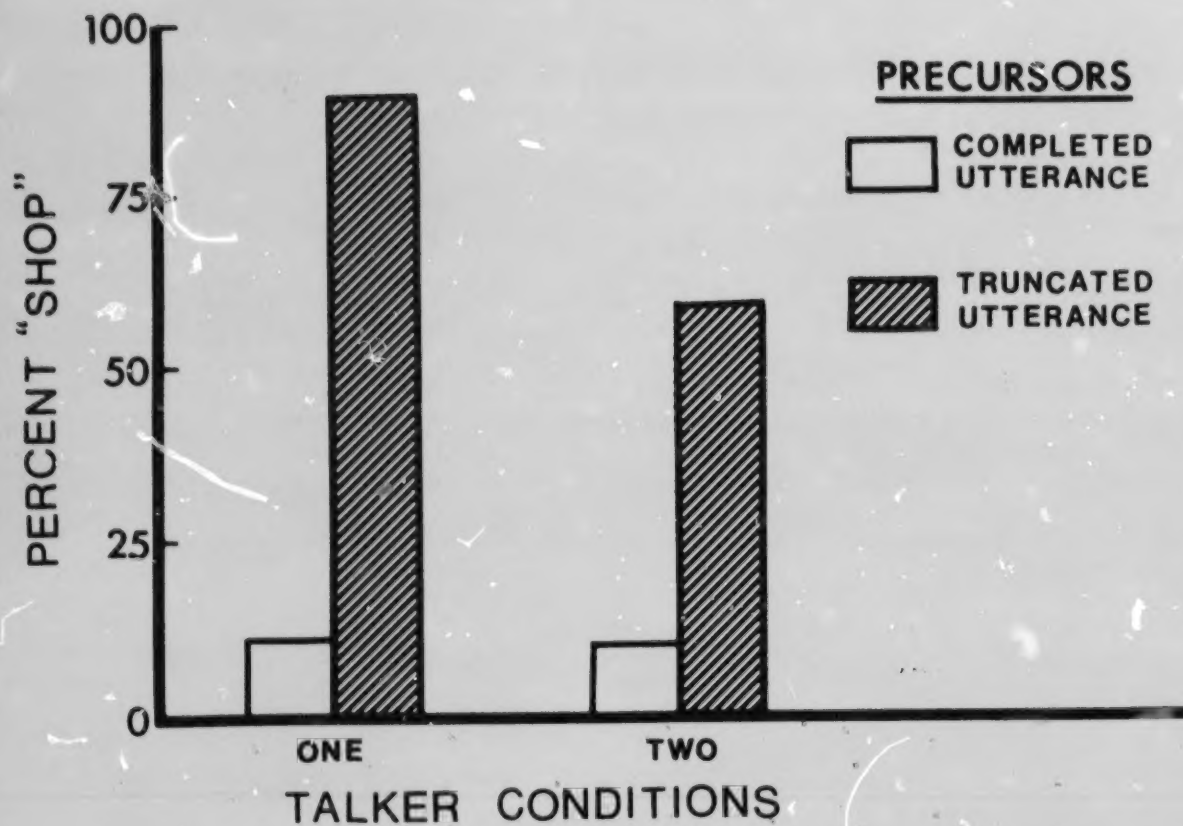


Figure 3. Percentage of shop responses for stimuli pairing different precursors with 0 msec of silence and the chop token.

talker manipulation had essentially no influence on shop tokens and a modest influence on chop tokens.

Discussion

In the present study, the precursor utterances contrasted in two ways. They contrasted first in terms of the manner of their production, the precursors either having been spoken as completed utterances (in fact, as completed sentences) that had not been articulated continuously with a following word, or having been spoken as ongoing utterances that had been articulated continuously with a following word. The precursors also contrasted in terms of their source, having been spoken either by the same person who had produced the target words or by a different person. The significance of silence for listeners' phonetic decisions was only weakly related to this latter source contrast. No main effect of talkers was observed for the shop stimuli and a comparatively small effect was observed for the chop stimuli.

The phonetic significance of silence was closely related to the contrast in the manner of production, however. Fricative-affricate judgments made following completed-utterance precursors were seldom influenced by the silence cue, while judgments made following truncated-utterance precursors were frequently influenced by silence. Thus, silence contributed to phonetic perception only in the context of ongoing articulation. This result can be accounted for in terms of the fact that silence is a reliable cue to closure only in the context of continuous articulation from a talker. Apparently, there was sufficient information in the truncated-utterance precursors to specify that they could have been articulated continuously with a target, and correspondingly, there was information in the completed-utterance precursors to specify that continuous articulation was not possible. Interestingly, this was the case both for the one-talker and for the two-talker stimuli, suggesting that the information for continuous articulation is, in some measure, independent of the information for vocal tract source.

It should be noted that our observation that silence may have phonetic significance in the context of some two-talker stimuli differs from the earlier finding of Dorman et al. (1979). Perhaps this is because several sources of suprasegmental continuity were emphasized in the current experiment. Both the male talker and the female talker recorded their utterances at the same rhythm, and this temporal constraint, coupled with the coordination of stress patterns across talkers, may have enhanced the suprasegmental similarity of their utterances. Also, the prosodic contrasts of the current experiment may have highlighted a perceptual dimension that did not serve to contrast the stimuli of the previous experiment. Even so, some effect of source discontinuity was demonstrated here; a main effect for number of talkers and an interaction of number of talkers and precursor type were both observed for the chop-token stimuli.

Two methodological points might be raised concerning this experiment. First, with regard to the observations about the effects of the manner of precursor production on phonetic perception, it might be noted that the boundary for a cue to any phonetic contrast may vary as a function of context. For example, Repp et al. (1978) report that in a rapid sentence context (as

defined by a rapidly articulated precursor) subjects require more silence to cue the perception of an affricate than is needed to cue the same affricate in a slow context. It is conceivable that the effects reported in the current study were obtained because the completed-utterance precursors and the truncated-utterance precursors specified different "boundaries" for the duration of silence needed to cue the affricate. If this is true, then one cannot conclude that the completed-utterance precursors disrupted the perceived continuity of stimuli; they merely shifted the boundary for the silence cue. The 75 msec silent interval used here may have been too brief for the shop token to be heard as chop following completed-utterance precursors. It is difficult to frame a comparable argument for the chop stimuli, since the silent interval presented was maximally brief (0 msec). However, at least for the shop stimuli, it is important to test whether the precursor differences observed in the current experiment were an artifact of the specific silent interval used. Experiment 2 repeats a portion of Experiment 1 with the addition of a range of silent intervals surrounding the 75 msec interval.

A second concern relates to the observation that the information for phonetic perception and the information for the source of an utterance are in some measure independent. It might be argued, in direct opposition, that while the number of talkers producing these utterances was varied, the number of talkers perceived to have produced them did not vary in corresponding fashion. If so, the phonetic significance of silence could remain strictly and exclusively tied to listeners' perceptions that an entire stimulus was produced by a single vocal tract source. By this hypothesis, responses to the two-talker truncated-utterance stimuli greatly resembled those to the one-talker truncated-utterance stimuli because listeners failed to detect the discontinuity in vocal tract source in the two-talker case. Also, silence might be said to have had little phonetic significance in all completed-utterance stimuli because listeners heard those stimuli as having been produced by two talkers. It is important, therefore, to determine how many sources were heard by listeners in each experimental condition. Experiment 3 repeats a portion of Experiment 1, with additional instructions to subjects to report the number of talkers heard, as well as the target word heard.

EXPERIMENT 2

In order to test the hypothesis that the precursor manipulation altered the duration of silence needed to cue the affricate (rather than the perceived continuity of the stimuli), this experiment assessed listeners' perceptions of the shop token when it followed either a completed-utterance or truncated-utterance precursor with a range of intervening silent intervals. Since all completed-utterance precursors yielded similar results in Experiment 1, only the completed-utterance statement was included here.

Method

Two sets of stimuli were generated for this experiment, one by pairing a range of silent intervals and the shop token with the male talker's completed-utterance precursor statement and the other by pairing them with his truncated-utterance precursor. Nine different silence durations were included in each stimulus set (0, 15, 30, 45, 60, 75, 90, 105, and 120 msec). Eight

repetitions of each of the 18 stimuli were randomized and presented to nine listeners, their task and test conditions being the same as those of Experiment 1.

Results and Discussion

In Figure 4 it can be seen that the shop-token was heard as chop for a relatively constant and small percentage of the completed-utterance stimuli at every silent interval. By contrast, the percentage of chops heard varied with the silent interval for truncated-utterance stimuli, and at the longer intervals nearly all stimuli were heard as chop. An analysis of variance showed this difference between the truncated-utterance and completed-utterance precursors to be highly significant, $F(8,56)=17.29$, $p<.001$.

The effect of precursor-utterance completion demonstrated in Experiment 1 cannot be accounted for simply by the fact that completed-utterance precursors alter the absolute duration of silence that is needed to cue the affricate. Following the completed-utterance precursor statement, listeners gave relatively few chop responses at every silent interval from 0 to 120 msec. Thus, there is no evidence for a boundary at silent intervals greater than 75 msec; in fact, there is no evidence for a phonetic boundary at all. This suggests that silence was not being interpreted phonetically and that the completed-utterance precursor was perceived to be discontinuous with the target. In contrast, following the truncated-utterance precursor, listeners' responses varied as a function of the silent interval. At the briefest silent interval (0 msec), fewer than 7% of the stimuli were perceived as chop, while at longer intervals (30 msec or greater) nearly all of the stimuli were perceived as chop. This is clear evidence for the phonetic significance of silence in the presence of the truncated-utterance precursor. In general, these results support the conclusion that listeners perceived the truncated-utterance precursor to have been produced continuously with the target word and perceived the completed-utterance precursor to have been discontinuous with the target.

EXPERIMENT 3

This experiment was designed to test the hypothesis that the phonetic significance of silence is solely contingent upon the perception that the entire utterance containing the silence is produced by a single talker. If this is the case, it would be expected that listeners' reports of the number of talkers heard to produce an utterance would be a good predictor of their phonetic responses to the silence cue in that utterance.

Method

The shop-token target and 75 msec of silence were paired with the male talker's completed-utterance precursor statement, the male's truncated-utterance precursor, the female talker's completed-utterance precursor, and the female's truncated-utterance precursor. Ten repetitions of each of these four stimuli were randomized and presented to eight listeners under the same testing conditions as those described in Experiment 1. The task instructions called for two responses: listeners were asked first to identify the number

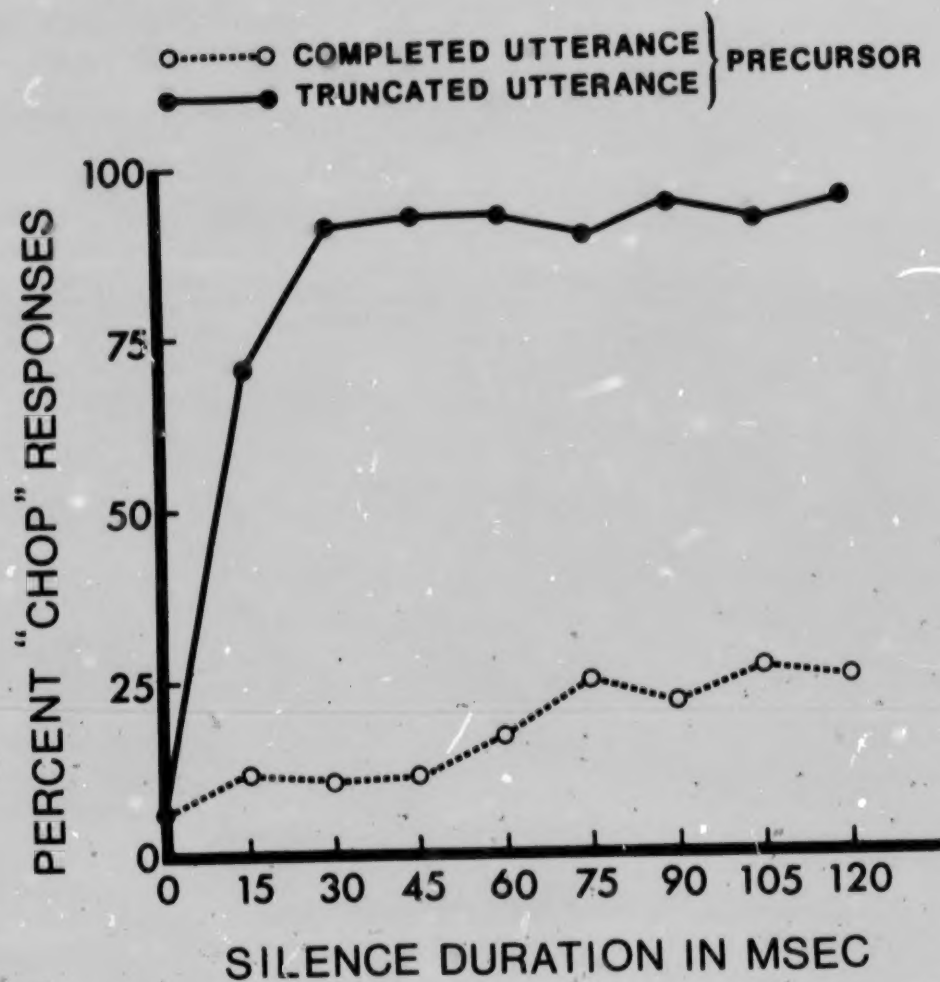


Figure 4. Percentage of chop responses for stimuli pairing different precursors with a range of silent intervals and the shop token.

of talkers heard to produce a stimulus as either one or two and then to identify the final word in each stimulus as either shop or chop.

Results and Discussion

Listeners correctly identified 81.2% of the one-talker completed-utterance stimuli and 96.2% of the one-talker truncated-utterance stimuli as having been produced by one talker. Their identification of the number of sources producing the two-talker stimuli again was quite good, as 95.0% of the completed-utterance stimuli and 96.2% of the truncated-utterance stimuli were identified as having been produced by two talkers.

Phonetic classifications were analyzed only for the stimuli that had been accurately classified according to the number of talkers heard. These results are summarized in Figure 5, and the pattern is similar to that of Experiment 1. The silence manipulation consistently influenced phonetic judgments following truncated-utterance precursors but not following completed-utterance precursors. As in the earlier experiment, this difference was highly significant, $F(1,7)=184.01$, $p<.001$. By contrast, phonetic judgments did not differ significantly between stimuli judged to have been produced by one talker and those judged to have been produced by two talkers, $F(1,7)=1.16$.

The suprasegmental information for utterance continuity manipulated in this experiment would appear to be largely independent of the information for source identity. Listeners were accurate in their identification of the number of talkers who produced these stimuli; however, their phonetic responses were more accurately predicted by the utterance completion manipulation than by the number of talkers identified. It would appear that the phonetic significance of silence cannot be related solely (or even principally) to these listeners' perceptions of the number of vocal tract sources producing an utterance.

GENERAL DISCUSSION

The results of the experiments reported here are consistent with a growing body of evidence that suggests that listeners integrate acoustic cues in speech perception with reference to constraints on speech production (Bastian et al., 1961; Repp et al., 1978; Dorman et al., 1979). In these experiments, silence influenced phonetic perception only when it followed precursors that had been produced continuously with a following word, and it is only in this circumstance that silence implicates vocal tract closure uniquely. When the surrounding speech events are produced separately, silence might be the result of an articulatory discontinuity, rather than the consequence of the closure that distinguishes articulation of an affricate from that of a fricative.

This articulatory argument concisely accounts for the one-talker data of the current study and it highlights the parallel to previous research. A comparable account of the two-talker data is less clear. The distinction between the completion and the truncation of precursor articulation is again applicable and predicts the data nicely. However, an articulatory rationale for the two-talker result is less persuasive given the detectability of the

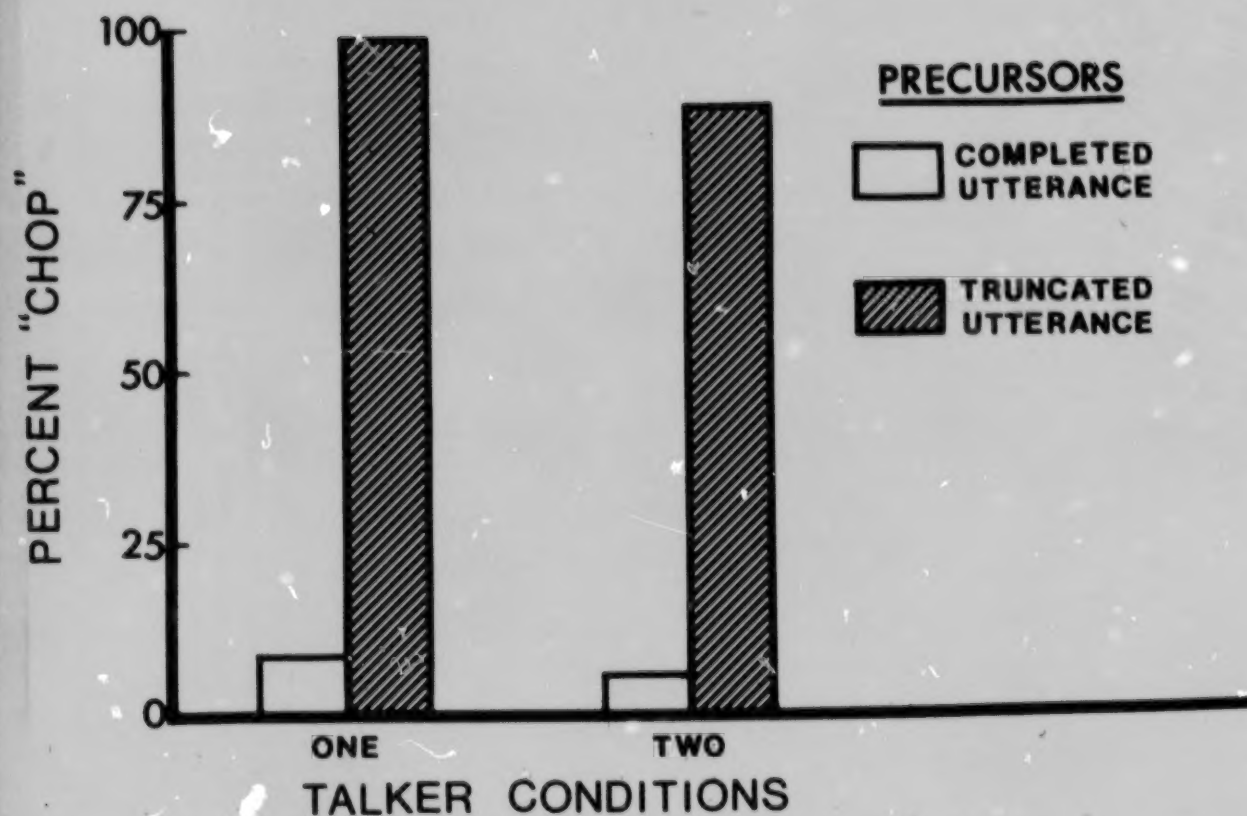


Figure 5. Percentage of chop responses for stimuli pairing different precursors with 75 msec of silence and the shop token, given that a listener had correctly identified the number of talkers who had produced these stimuli.

change in articulatory source. One possibility is to think of these stimuli as providing the listener with conflicting information about articulation. From this perspective, one could argue that the suprasegmental information for ongoing articulation generally was more powerful than the source information for a break in articulation. This should not be altogether surprising given the efforts to match stress-timing and prosody in the two talkers' productions of the original utterances.

In another sense, however, the two-talker result might be taken as strong evidence that a source-independent articulatory coherence underlies the perception of speech. When speech is characterized not just as a sequence of vocal tract configurations cued by source-specific acoustic features, but as a patterning of gestures arrayed in time, then it becomes possible to argue that two talkers could adopt comparable articulatory strategies (and produce comparable gestural patterns) despite differences in their vocal tract dimensions. In other words, speech might be defined as a particular organization over the articulatory system; this organization could be common to all speakers of a dialect independent of specific values on such source properties as vocal tract length or glottal frequency. If, for the purposes of phonetic perception, listeners were principally sensitive to the acoustic patterning that reflects this organization, then the critical factor for perception would be information about patterning at the suprasegmental level. A change in vocal tract source would not be a competing factor, since idiosyncratic source properties do not enter into the definition of articulatory patterning. In short, one could argue that, for purposes of phonetic perception, listeners are sensitive to information that is largely orthogonal to source properties; as a result, they are largely insensitive to changes in those properties when information for articulatory patterning is preserved. This argument would account for our findings in both the one-talker and two-talker conditions.

In sum, the effects observed here may be especially interesting for several reasons. First, they support the view that the speech perceiver integrates acoustic cues with reference to speech articulation; silence is phonetically significant because it is a cue to vocal tract closure. Second, they suggest that silence supplies phonetic information only when it is perceived to have occurred in an ongoing articulation, and finally, there is evidence that the information that is relevant for phonetic perception is, to some degree, independent of source.

REFERENCES

- Bastian, J., Eimas, P., & Liberman, A. M. Identification and discrimination of a phonemic contrast induced by silent interval. Journal of the Acoustical Society of America, 1961, 33, 842.
- Boomer, D. S., & Laver, J. D. Slips of the tongue. The British Journal of Disorders of Communication, 1968, 3, 2-12.
- Cooper, W. E., Lapointe, S. G., & Paccia, J. M. Syntactic blocking of phonological rules in speech production. Journal of the Acoustical Society of America, 1977, 61, 1314-1320.

- Cooper, W. E., Egido, C., & Paccia, J. M. Grammatical control of a phonological rule: palatalization. Journal of Experimental Psychology: Human Perception and Performance, 1978, 2, 264-272.
- Dorman, M. F., Raphael, L. J., & Liberman, A. M. Some experiments on the sound of silence in phonetic perception. Journal of the Acoustical Society of America, 1979, 65, 1518-1532.
- Gerstman, L. Cues for distinguishing among fricatives, affricates, and stop consonants. Unpublished doctoral dissertation, New York University, 1957.
- Lieberman, P. Intonation, perception, and language. Cambridge, Mass.: M.I.T. Press, 1967.
- Repp, B. H., Liberman, A. M., Eccardt, T., & Pesetsky, D. Perceptual integration of acoustic cues for stop, fricative, and affricate manner. Journal of Experimental Psychology: Human Perception and Performance, 1978, 4, 621-637.

LARYNGEAL ADJUSTMENTS IN STUTTERING: A GLOTTOGRAPHIC OBSERVATION USING A MODIFIED REACTION PARADIGM*

Hirohide Yoshioka* and Anders Lofqvist**

Abstract. An experimental paradigm for studying stuttering is described. Simultaneous recordings of the voice and a photo-electric glottogram were made during speech production in a modified reaction time situation. Preliminary results of one subject revealed that some cases of dysfluency (word initial voiceless stop repetitions) were accompanied by one or two extra glottal opening gestures corresponding to repeated release bursts. Similar "trial-and-error" types of glottal movements were also detected during silent delayed response time in other productions that were perceptually judged as fluent. As for the production of words beginning with voiced sounds, unnecessary opening gestures prior to the vocalization for the initial voiced segment were sometimes observed for both fluent and dysfluent utterances. These findings suggest that stuttering is linked to a temporal disruption of the control of the abductory and adductory gestures of the glottis relative to supraglottal articulation and respiratory function.

INTRODUCTION

Although its ultimate etiology still remains unclear, stuttering has been extensively investigated from various viewpoints, including acoustical and physiological examination (e.g., Adams, 1978). Various types of modified reaction time paradigms using visual, auditory, and tactile stimuli have been frequently used to study not only verbal but also non-speech behavior in stutterers (e.g., Adams & Hayden, 1976; Starkweather, Hirschman, & Tannenbaum, 1976; McFarlane & Prins, 1978). On the whole, statistical analyses in these studies indicate a general tendency towards longer latency for stutterers than for normal subjects, although the magnitude of the difference reported has varied. It is worth noting that even perceptually "fluent" productions by stutterers have been sometimes found to be accompanied by longer reaction time than for normals.

*A version of this paper was presented at the 98th Meeting of the Acoustical Society of America in Salt Lake City, Utah, 26-30 November 1979.

+Also University of Tokyo, Tokyo, Japan.

**Also Lund University, Lund, Sweden.

Acknowledgment: This work was supported in part by NINCDS Grants NS-13617 and NS-13870, and Grant RR-05596 to Haskins Laboratories. We thank Katherine S. Harris for her encouragement of this project and her comments on an earlier version of this paper. We are also grateful to Patrick W. Nye, Thomas Baer, Nancy S. McGarr and Carole E. Gelfer for their cooperation in conducting the experiment.

These studies, however, do not allow us detailed information on the dynamics of the production of fluent and dysfluent utterances, since the method yields only a single datum, i.e., the latency of the acoustic signal for each production. It might be more revealing to explore aspects of the underlying temporal organization of articulation, particularly during the time lag from a trigger signal to the output explicitly related to speech performance. More specifically, does the stutterer's apparently "fluent" production, when it is accompanied by an abnormally long reaction time, reflect a delay in the initiation of activity of all components of the system, or is there a disruption of the coordination of parts of the system? One of the aims of this study is to examine the acoustic signal in parallel with observation of the glottal opening and closing gestures as a function of time in response to visual stimuli, using the photo-electric glottographic technique in parallel with endoscopic monitoring through a laryngeal fibroscope.

Another current trend in experimental studies of stuttering focuses on phonetic structures such as the voicing feature and its relationship to occurrences of dysfluency (Wingate, 1969; Adams & Reis, 1971; Adams & Reis, 1974; Wall, Starkweather, & Harris, submitted). Furthermore, the initiation of an utterance is often reported to be particularly difficult for stutterers, in that high incidences of dysfluency are found in this position (e.g., Quarrington, Conway, & Siegel, 1962). Therefore, we paid special attention to the initial segments of isolated meaningful words particularly with regard to the voicing feature.

METHOD AND PROCEDURE

Table 1 shows the word list used in the experiment, which contains many types of initial phonemes and various combinations of phonemes. These words were chosen, because they have a simple phonetic structure and are familiar to native speakers of English. Note that there are several minimal pairs with respect to the voicing of initial obstruents, consisting of /CV/ and /CVC/. Also included are a few words beginning with a vowel or a liquid. Only high front vowels /i/, /I/, /e/ and /æ/ were used, since the tongue movement for back vowel production can interfere with the passage of light in photo-electric glottography.

Table 1

List of test words

1. pea	10. bit	19. thief	28. they
2. peak	11. bay	20. thick	29. ray
3. peach	12. bisk	21. zee	30. risk
4. pit	13. sea	22. zip	31. lay
5. pay	14. seek	23. jape	32. lisp
6. pip	15. sick	24. jest	33. eat
7. bee	16. shape	25. jet	34. it
8. beak	17. sheep	26. jig	35. eight
9. beach	18. ship	27. this	36. ache

Figure 1 presents the experimental procedure. The subject was seated in front of a screen, on which one of the 36 randomized words appeared in capital letters under computer control. After a two second display of each word, the word disappeared for a variable period of one to five seconds, and was followed by a "start" signal, represented by four stars on the screen. The subject was requested to pronounce the specified word as quickly as possible in response to the signal. Two timing marks, which corresponded to the onset of the word and "start" signal presentations, were recorded on instrumentation tape, together with acoustic recordings of the utterances. The photo-electric glottographic signals, sensed by a photo-transistor placed on the neck just below the lower edge of the cricoid cartilage, were recorded on a third channel. The fiberscope, used as the light guide, enabled continuous monitoring of the glottal image through the viewfinder during the entire session. This combined technique, we believe, permits reliable interpretation of the large amounts of photo-electric recordings during glottal abduction and adduction movements (Yoshioka, Löfqvist, & Hirose, 1979; Löfqvist & Yoshioka, 1979). Figure 1 contains a sample of the simultaneously recorded signals using a normal subject. In this case, the presentation of the word "sick" was followed by a three second delay of the "start" signal display.

The subject of this experiment was referred from another study of stuttering. He was a college-age male, with no diagnosed abnormalities of speech or hearing other than a history as a stutterer. While he did not produce many obvious non-fluencies of speech in the experimental condition, he described himself as a moderate stutterer.

Before the actual recordings of these parameters, the subject was allowed to practice by use of a "dummy" list, consisting of ten similar words (not shown here) under the same experimental conditions. Then the word list (shown in Table 1) was presented for a minute to the subject for inspection. In each session, the recordings were repeated at least three times using the list in random order.

RESULTS

In this preliminary report, we will focus on examples of interest from the single subject.

Figure 2 shows the glottographic curve and the audio envelope for a production of the word "pay" by the subject, which was perceptually judged as dysfluent. The vertical line corresponds to the onset of the start signal display. This graph shows that repetitions of word initial voiceless stop bursts were accompanied by glottal opening gestures corresponding to each repeated burst and aspiration. A closer look further reveals that the first oral release occurred just after the glottis had closed, the second just before glottal closure, and the third during glottal opening. In other words, the first two oral releases were improperly timed with respect to the glottal movements, at least to produce a sufficient aspiration noise for a word initial voiceless stop in American English. Moreover, the relative timings of oral and laryngeal articulations are gradually brought into the proper relationship during the repetitions. These findings also imply that the repetitions of the voiceless stop in this case are associated with a temporal disruption of the coordination between oral and laryngeal articulations.

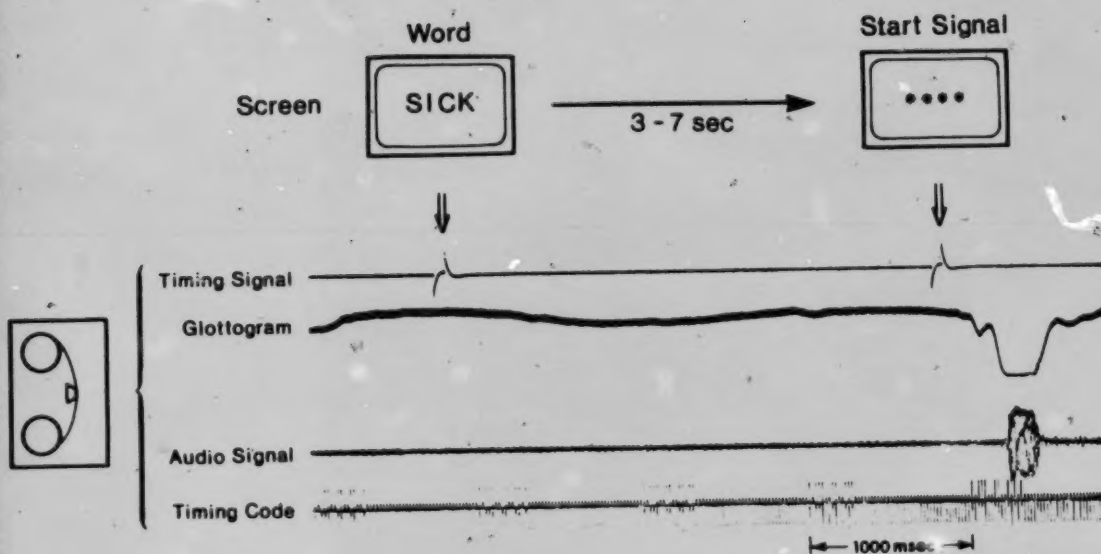


Figure 1. Diagram of experimental procedure.

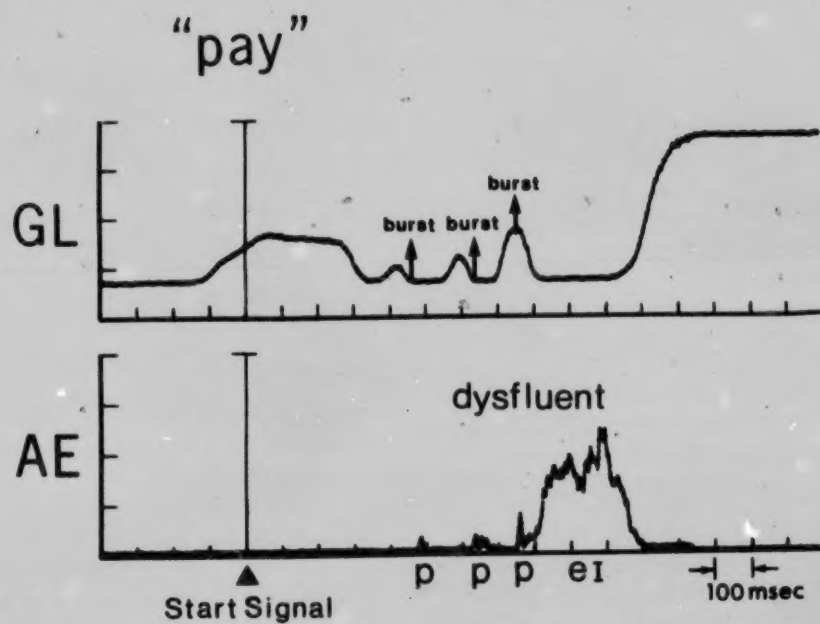


Figure 2. Glottogram and audio envelope for production of the word "pay."

Figure 3 presents a production of the word "bee" by the same stutterer. Although the utterance was fluent, the acoustic latency of more than 900 msec is far longer than normal for the subject. Moreover, two presumably unnecessary opening gestures of the glottis well before the initial voiced sound production can be found. This finding may be interpreted as evidence that even the fluent speech by a stutterer can be "dysfluent," in that inappropriate articulations can be made, even though no acoustic signal is generated.

Figure 4 shows production of the word "ship," which was judged as fluent. In this case, the onset of frication noise for the initial voiceless fricative almost coincided with the second peak of the two opening gestures. Since normal subjects seldom if ever show two separate glottal opening gestures for an initial single voiceless fricative in English (e.g., Yoshioka, Löfqvist, & Hirose, 1979), it may be that the first gesture was accompanied by an insufficient supply of air from the respiratory system to generate an audible frication noise. More specifically, it is possible in this case that a timing breakdown of the respiratory and laryngeal components occurred, despite the clinical judgment of a fluent production.

As for the production of words beginning with vowels, Figure 5 shows the word "it." The glottis started to close about 100 msec after the "start" signal, and then at least three repetitions of the initial vowel production were auditorily detected. (Although an additional noise, marked as "?" in the graph, was noticed in the audio waveform as well as in the audio envelope, sound spectrograms did not show any formant structure, and it seemed rather to be a mechanical noise. Nevertheless, the possibility that the initial noise was another stuttered [I] production cannot be completely ruled out.)

Figure 6 is another production of the word "it" by the same stutterer. In this case, the glottis had been closed after about 250 msec, although the initial stuttered [I] did not occur until approximately 200 msec later. Therefore, a plausible reason why the initial sound was not produced after a presumably appropriate glottal closure had been attained is again a disruption of the timing of laryngeal and respiratory functions.

Figure 7 contains the word "eight," which was judged as fluent, but the audio envelope as well as sound spectrograms clearly show two bursts for the word final /t/ production. Besides the /t/ problem, the long delay between the glottal closure and the acoustic output for the fluent [eI] production is quite obvious. This long latency may again be related to the timing discordance of laryngeal and respiratory events. Admittedly, the causal relationship still remains unclear: One interpretation is that the air supply from the lungs was simply delayed. An alternative view is that the laryngeal closure was too tight to permit voicing during the long reaction time. Since photo-electric glottography only allows us registration of glottal opening, variations during laryngeal closure without sound production (e.g., the degree of glottal closure "tightness" or the presence of false cord approximation) cannot be explicitly explored by this method.

Finally, Figure 8 shows a production of the word "eat." Although it was judged as dysfluent in that repeated [I] sounds were heard, the glottogram also revealed a paradoxical opening gesture prior to the vocalization for the initial vowel segment. It should be emphasized that such apparently abnormal

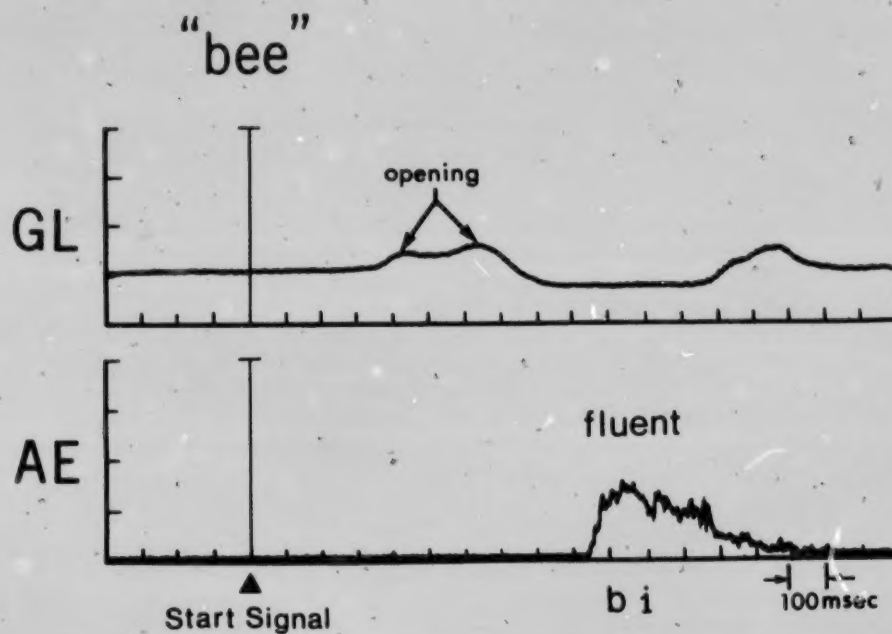


Figure 3. Glottogram and audio envelope for production of the word "bee."

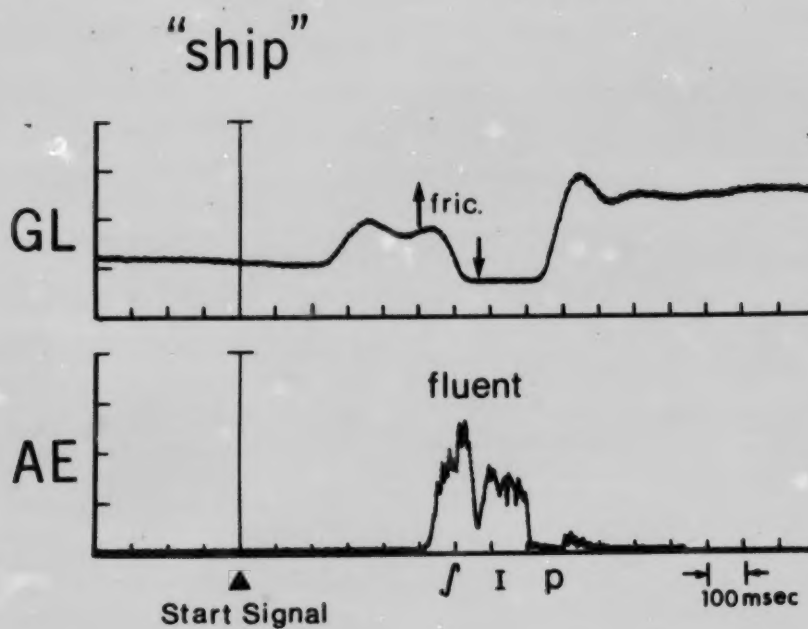


Figure 4. Glottogram and audio envelope for production of the word "ship."

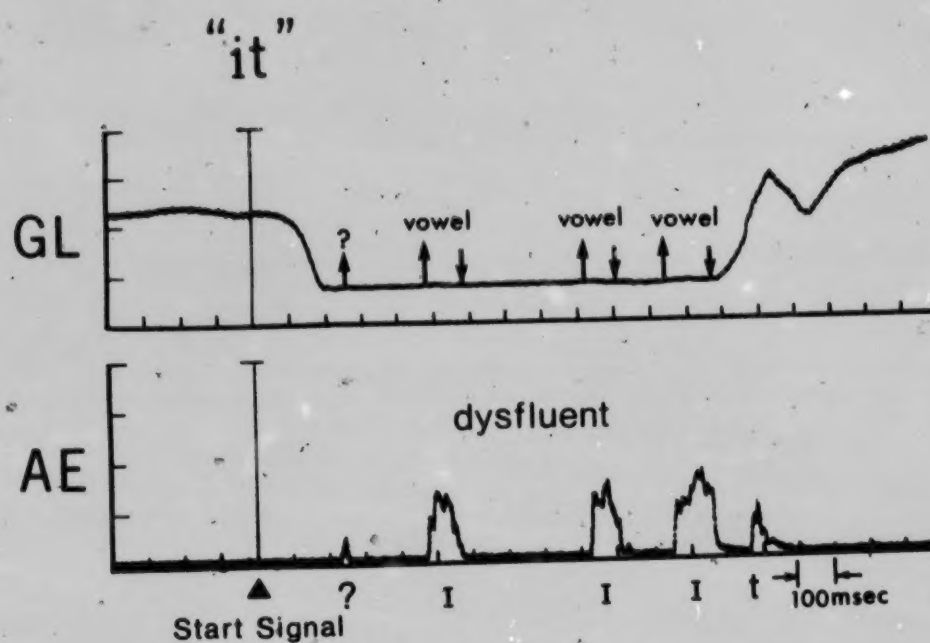


Figure 5. Glottogram and audio envelope for production of the word "it."

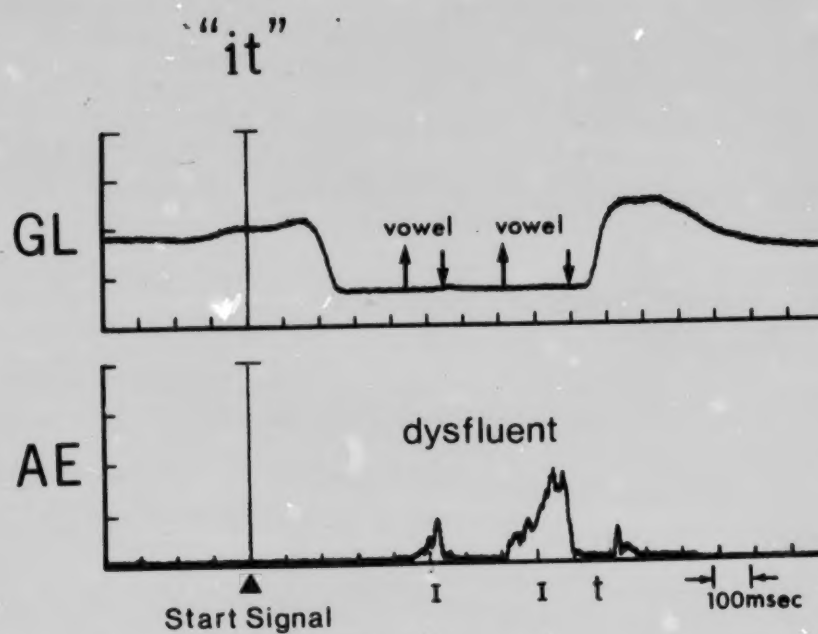


Figure 6. Glottogram and audio envelope for another production of the word "it."

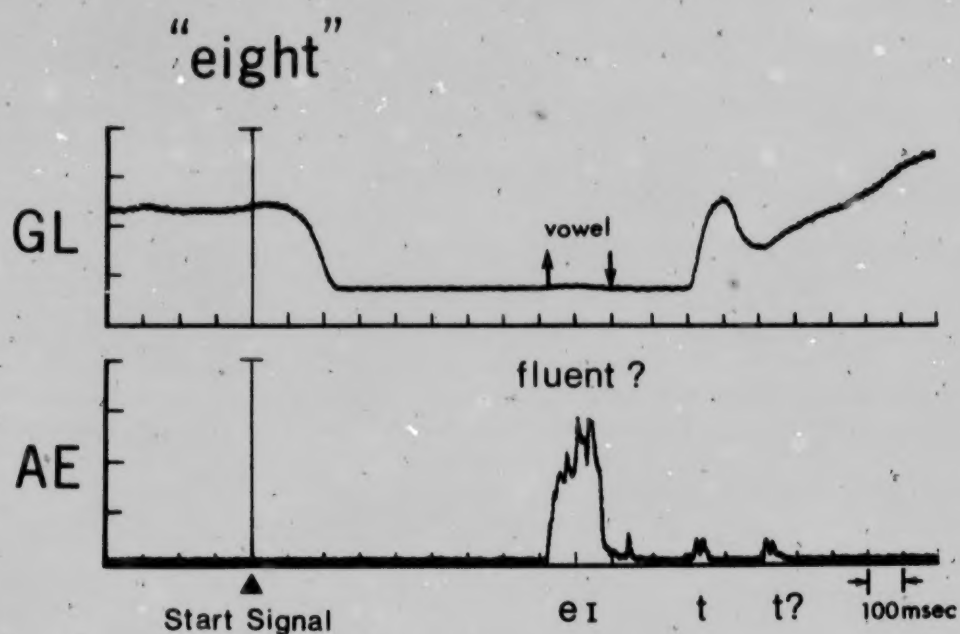


Figure 7. Glottogram and audio envelope for production of the word "eight."

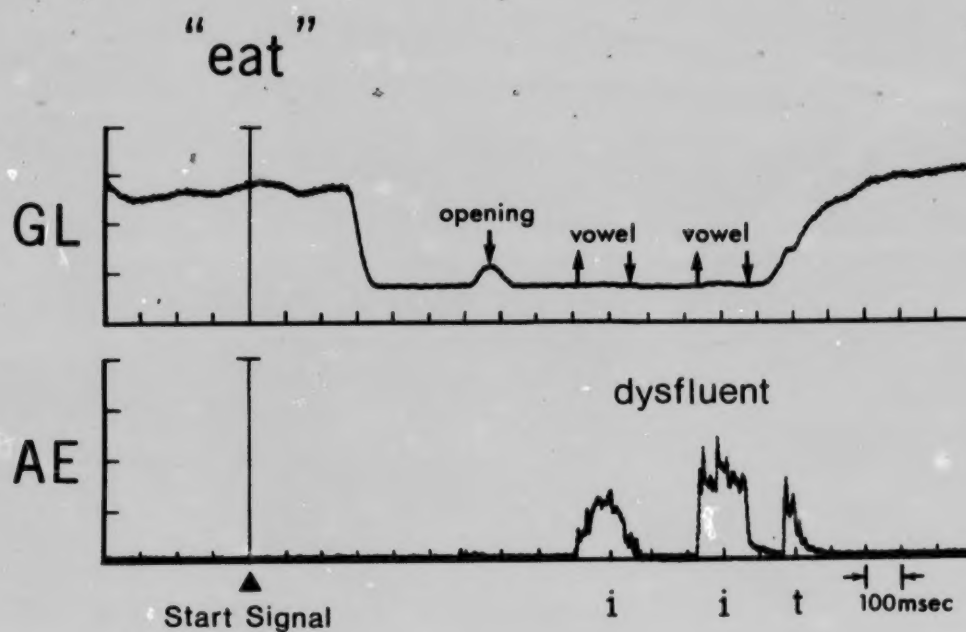


Figure 8. Glottogram and audio envelope for production of the word "eat."

movements can occur, even when they are not directly related to the perceptual judgment of fluency.

In summary, the experiment described above illustrates the feasibility of applying the transillumination technique, in a modified reaction time paradigm, to the study of stuttering. The examples here show that stuttering may be linked to a temporal disruption of the control of the abductory and adductory gestures of the glottis, particularly in relation to supraglottal articulation and respiratory function in speech. In addition, it is suggested that even perceptually fluent utterances by a stutterer can be associated with underlying abnormal articulations, regardless of their direct relationship to the acoustic output and therefore to the clinical judgment of fluency. We are currently collecting and analyzing more data on various stutterers, differing according to "types" of dysfluency (repetitions, prolongations, and blocks), primary locus of "tension" (labial, lingual, or laryngeal) and "severity" (mild, moderate, or severe). These results, including a combined monitoring of respiratory function, will be reported elsewhere in the near future.

REFERENCES

- Adams, M. R. Stuttering theory, research and therapy: The present and future. Journal of Fluency Disorders, 1978, 3, 139-148.
- Adams, M. R., & Hayden, P. The ability of stutterers and nonstutterers to initiate and terminate phonation during production of an isolated vowel. Journal of Speech and Hearing Research, 1976, 19, 290-296.
- Adams, M. R., & Reis, R. The influence of the onset of phonation on the frequency of stuttering. Journal of Speech and Hearing Research, 1971, 14, 639-644.
- Adams, M. R., & Reis, R. Influence of the onset of phonation on the frequency of stuttering: A replication and reevaluation. Journal of Speech and Hearing Research, 1974, 17, 752-754.
- Löfqvist, A., & Yoshioka, H. Laryngeal activity in Swedish voiceless obstruent clusters. Haskins Laboratories Status Report on Speech Research, 1979, SR-59/60, 103-124.
- McFarlane, S. C., & Prins, D. Neural response time of stutterers and nonstutterers in selected oral motor tasks. Journal of Speech and Hearing Research, 1978, 21, 768-778.
- Quarrington, B., Conway, J. K., & Siegel, N. An experimental study of some properties of stuttered words. Journal of Speech and Hearing Research, 1962, 5, 388-394.
- Starkweather, C. W., Hirschman, P., & Tannenbaum, R. S. Latency of vocalization onset: Stutterers versus nonstutterers. Journal of Speech and Hearing Research, 1976, 19, 475-480.
- Wall, M. J., Starkweather, C. W., & Harris, K. S. The influence of voicing adjustments on the location of stuttering in the spontaneous speech of young children. Submitted to Journal of Speech and Hearing Research.
- Wingate, M. E. Sound pattern in artificial fluency. Journal of Speech and Hearing Research, 1969, 12, 677-686.
- Yoshioka, H., Löfqvist, A., & Hirose, H. Laryngeal adjustments in the production of consonant clusters and geminates in American English. Haskins Laboratories Status Report on Speech Research, 1979, SR-59/60, 127-151.

MISSING -ING IN READING: LETTER DETECTION ERRORS ON WORD ENDINGS*

Adam Drewnowski+ and Alice F. Healy++

Abstract. Four experiments in which subjects searched for the letter n yielded many detection errors on the suffix morpheme -ing. The subjects made more errors on word endings than in the middle of the words and more errors on -ing than on -en, -ion, or -ment endings. More errors were made when the -ing ending was a suffix morpheme than when it formed part of the word stem. These results suggest that the combination of the high orthographic frequency of the letter sequence -ing, its spatial predictability, and its function as a suffix morpheme lead it to serve as a single unit in reading.

In previous studies we have shown that subjects searching for instances of a given target letter in printed text make most detection errors on the very common function words the and and (Drewnowski & Healy, 1977; Healy, 1976). The subjects' tendency to make errors on the word the seems to be a function of reading skill, since the pattern of errors made by school children was directly related to their reading ability (Drewnowski, 1978). The present letter detection paradigm (Corcoran, 1966; Healy, 1976) may therefore reflect reading processes better than detection tasks used by other investigators (e.g., Johnson, 1975; Wheeler, 1970), which do not involve searching for targets in paragraphs of text. We now use the letter detection paradigm to examine the processes involved in reading multisyllabic words. Specifically, we focus on the pattern of errors that are made on frequent word endings. It has been suggested that frequent word endings and common function words can both serve as word boundary morphemes (Kean, 1977). Our hypothesis is that this linguistic similarity will be reflected in the reading task.

We have previously interpreted the disproportionate number of letter-detection errors on the most common words in terms of reading units of different sizes. Healy (1976, 1980) demonstrated more errors on common than on rare nouns and therefore proposed that the "missing letter effect" was caused by the subjects' tendency to process common words in units larger than the letter. Furthermore, the finding of more errors on function words

*Also in Journal of Verbal Learning and Verbal Behavior, 1980, 19, 247-262.

+University of Toronto

++Also Yale University

Acknowledgment. This research was supported by Grant APA 146 from the National Research Council of Canada, NICHD Grant HD01994, BRS Grant RR05596, NSF Grant BNS77-00077, and a grant from the Spencer Foundation. We are indebted to J. Chung for help with the data analyses of Experiment 4 and to W. K. Estes, T. K. Landauer, and A. G. Levitt for helpful discussions about this research.

occurring in familiar contexts than in inappropriate contexts may indicate that subjects process very familiar phrases in units that include more than one word (Drewnowski & Healy, 1977). The postulated reading units may be either perceptual (visual) or response (phonetic) units. The available data do not allow us to choose between these two possibilities. However, the data do provide evidence against alternative explanations of these results (see e.g., Corcoran, 1966), which refer to the syntactic and semantic redundancy of the words containing the target letters or the atypical pronunciation of the target letters (see Healy, 1976, 1980).

We now investigate the formation of reading units larger than the letter but smaller than the word, and we examine the conditions under which a given sub-word sequence of letters can serve as a single unit in the course of fluent reading. Whereas previously we were mainly concerned with whole-word frequencies (Thorndike & Lorge, 1944), we now distinguish between two types of sub-word frequency that may contribute to the formation of reading units. The first type, which we shall call orthographic frequency, is suggested by the work of Gibson (1965) showing that letter sequences forming frequent spelling patterns may in themselves function as higher order reading units. Central to this notion is the assumption that through repeated exposure to printed words, readers acquire in memory a representation of the type of letter sequences that can be expected to occur in English texts. It seems reasonable to assume that the most common letter sequences in English words will be most likely to be unitized and will thus most successfully "conceal" the identity of their constituent letters. As a result, we might expect most letter detection errors to occur on letter sequences that are orthographically most frequent.

The second type of frequency, which we shall call positional frequency, is suggested by the work of Mason (1975) and Katz (1977) showing that skilled readers are sensitive to the positional frequencies of single letters within test words. Extending this notion from single letters to letter sequences, we might expect that most letter detection errors will occur in letter sequences appearing in their most spatially redundant locations within the word.

EXPERIMENT 1

In order to discriminate between the effects of orthographic and positional frequency, we examined the incidence of detection errors on the target letter n embedded in two different word endings: -ing and -ent. Whole-word frequency, as established by Kučera and Francis (1967), was kept approximately constant.

The two letter sequences -ing and -ent were selected following a comparison of the Underwood and Schulz (1960) and the Mayzner, Tresselt, and Wolin (1965) frequency tables. Underwood and Schulz (1960) tabulated digram and trigram counts based on a random sample from the Thorndike and Lorge (1944) word list and on two additional samples of running text. According to their data, the embedded frequency of the trigram -ent (1778) exceeds that of the trigram -ing (1673).

Although this analysis provides a means of comparing the orthographic frequency of letter sequences, it does not take into account the position of

the letter sequences within the word. Such data are supplied by Mayzner et al. (1965), who tabulated trigram counts by word position and by word length for words from 3 to 7 characters in length. The trigram -ing was found to occur almost exclusively at the end of words, whereas the trigram -ent was less spatially predictable, occurring at a number of locations within the word.

The trigrams -ing and -ent thus differ in their orthographic and positional frequencies. If frequent letter sequences are perceived as higher order units (Gibson, 1965), then we might expect more letter detection errors to occur on words with -ent endings than with -ing endings. On the other hand, if positional frequency (Mason, 1975) is the critical factor in unit formation, then we might expect more errors to occur on the more spatially predictable -ing endings than on the -ent endings. The present design rules out single letter location or the word length in syllables as potential contributing variables, because the target letter n always occurred in the penultimate position within the word, and the length in syllables of test words was equated across both types of endings.

We further attempted to influence the number of letter detection errors by altering the nature of the search passage. In addition to a standard case prose passage, we employed a mixed case prose passage, in which alternating letters of the prose text were typed in upper and lower case. We also employed a meaningless passage of scrambled text. Such manipulations, designed to impede the formation of reading units larger than the letter or larger than the word, reduced the percentage of letter detection errors on the function words the and and in our previous studies (Drewnowski & Healy, 1977; Healy, 1980). We now wish to determine whether the unitization of sub-word letter sequences would also be affected by such procedures.

Method

Subjects. Fifty-four male and female students at the University of Toronto served as volunteer subjects in a group experiment conducted in the classroom.

Design and materials. Three 175-word passages, typed on separate sheets of paper, were constructed for the present experiment. Each passage included 42 "test" words (words that contain the target letter) and 133 "filler" words (words that do not contain any targets). The first passage, hereafter referred to as the "prose standard case passage," contained 18 test content words ending in -ing, 18 test content words ending in -ent, and 3 instances each of the function words in and and. The nature of the test content words was manipulated further by using 12 mono-, 12 di-, and 12 trisyllabic words, with half the words in each subgroup ending in -ing, and half ending in -ent. The target letter n thus always occurred in a function word or as part of a word ending.

In the case of monosyllabic words, both -ing and -ent trigrams necessarily formed part of the word stem (e.g., wing, went). In contrast, in trisyllabic words, -ing trigrams were always inflectional suffixes (e.g., applying) and -ent trigrams were most often derivational suffixes (e.g., persistent). The difference in nature of the test trigrams (word stem or

suffix) was systematically manipulated for disyllabic test words ending in -ing, with three words including -ing as part of the word stem (during, something, sterling) and three including -ing as a suffix (having, being, going). In addition, word stress was systematically manipulated for disyllabic words ending in -ent: Three of these took primary stress on the first syllable (recent, present, payment), and three took stress on the second syllable (resent, percent, torment). Since whole-word frequency was found to affect the incidence of letter detection errors (Healy, 1976, Experiment 4), we matched word frequencies (Kučera & Francis, 1967) across conditions wherever possible: The mean frequency of -ing test words was higher (181.1) than the mean frequency of -ent test words (122.7). However, trisyllabic words ending in -ing were specifically selected so that their mean frequency of usage (30.0) would be lower than the mean frequency of usage of trisyllabic words ending in -ent (93.3).

The second passage, hereafter referred to as the "prose mixed case passage," was identical to the prose standard case passage, except that alternating letters were typed in upper and lower case. Only one version of the prose mixed case passage was prepared, so the incidence of upper and lower case ns was randomized across targets.

The third passage, hereafter referred to as the "scrambled-word passage," was derived from the prose passages. The 36 test words and the three instances each of the function words in and and remained in the same locations as in the prose passages, but the order of the 133 filler words was randomized. No particular effort was made to place every test word in an inappropriate syntactic context: In fact, wherever two test words occurred in adjacent positions in the prose passages (e.g., "resent having" or "and rent"), their locations were preserved in the scrambled-word passage.²

Each passage was typed on a separate sheet of paper. The three passages arranged in all six possible sequences and preceded by a page of instructions were stapled together into booklets. The booklets were distributed according to a fixed rotation so that passage order was approximately counterbalanced across subjects.

Procedure. The subjects were instructed to read each passage at their normal reading speed, and to circle each instance of the target letter n. The subjects were told that if they ever realized that they had missed a target, they should not retrace their steps to encircle it. They were also told that they were not expected to detect all the ns so they should not slow down their reading speed in order to be overcautious about encircling the ns. The subjects were told to read the passages in the order in which they were stapled together and to go on to the next passage as soon as they had finished the preceding one.

Results

The results are summarized in Table 1, which includes for the three passages means and standard errors of the means for error percentages on the function words in and and and error percentages on the content words ending in -ing or -ent.

Table 1

Means (and Standard Errors) for Error Percentages in Experiment 1
as a Function of Passage Type and Word Type

Passage type	Word type			
	Function		Content	
	<u>in</u>	<u>and</u>	<u>ing</u>	<u>ent</u>
Prose standard case	77.0 (4.3)	80.7 (3.7)	33.5 (3.8)	15.1 (2.1)
Prose mixed case	78.3 (2.5)	53.7 (5.0)	24.5 (3.0)	7.0 (1.5)
Scrambled-word	72.0 (4.3)	75.6 (4.3)	37.1 (3.7)	13.2 (1.7)

Function words. Higher error percentages were obtained for the two function words in and and than for content words ending in -ing or -ent. The present data are thus consistent with our earlier reports (Healy, 1976, 1980; Drewnowski & Healy, 1977) that a disproportionate number of letter-detection errors occurs on frequent function words. In addition, the data preclude a purely phonological interpretation of the phenomenon, such as has been suggested by Corcoran (1966) and Locke (1978): Let us suppose that subjects are more likely to detect the letter n in words that contain the sound /n/, since this is the most common pronunciation of the letter n. One might then expect that an n in the sequence -ing (pronounced as /ŋ/) would be more difficult to detect than an n in a sequence that contains the conventional /n/ sound, such as in and and. However, considerably more errors were made on in and and than on the words containing -ing.

There were no differences in the percentage of errors on function words between the prose standard case and the scrambled-word passages [$t(106) = 1.01$, $p > .10$], possibly because the appropriateness of between-word context was not strictly controlled in the scrambled-word passage. Only one version of the prose mixed case passage was used such that all three instances of and appeared as aNd and all three instances of in appeared as In. Only the error frequency on the word and [$t(106) = 4.35$, $p < .01$] was reduced for the prose mixed case passage relative to the prose standard case passage. We interpret this finding to indicate that the word In was still perceived as a single unit by the readers, whereas the word aNd no longer was.

Content words. Table 1 also shows that more errors were made on words ending in -ing than on words ending in -ent, and that more errors were made on the prose standard case and scrambled-word passages than on the prose mixed case passage. An analysis of variance of error scores on content words showed

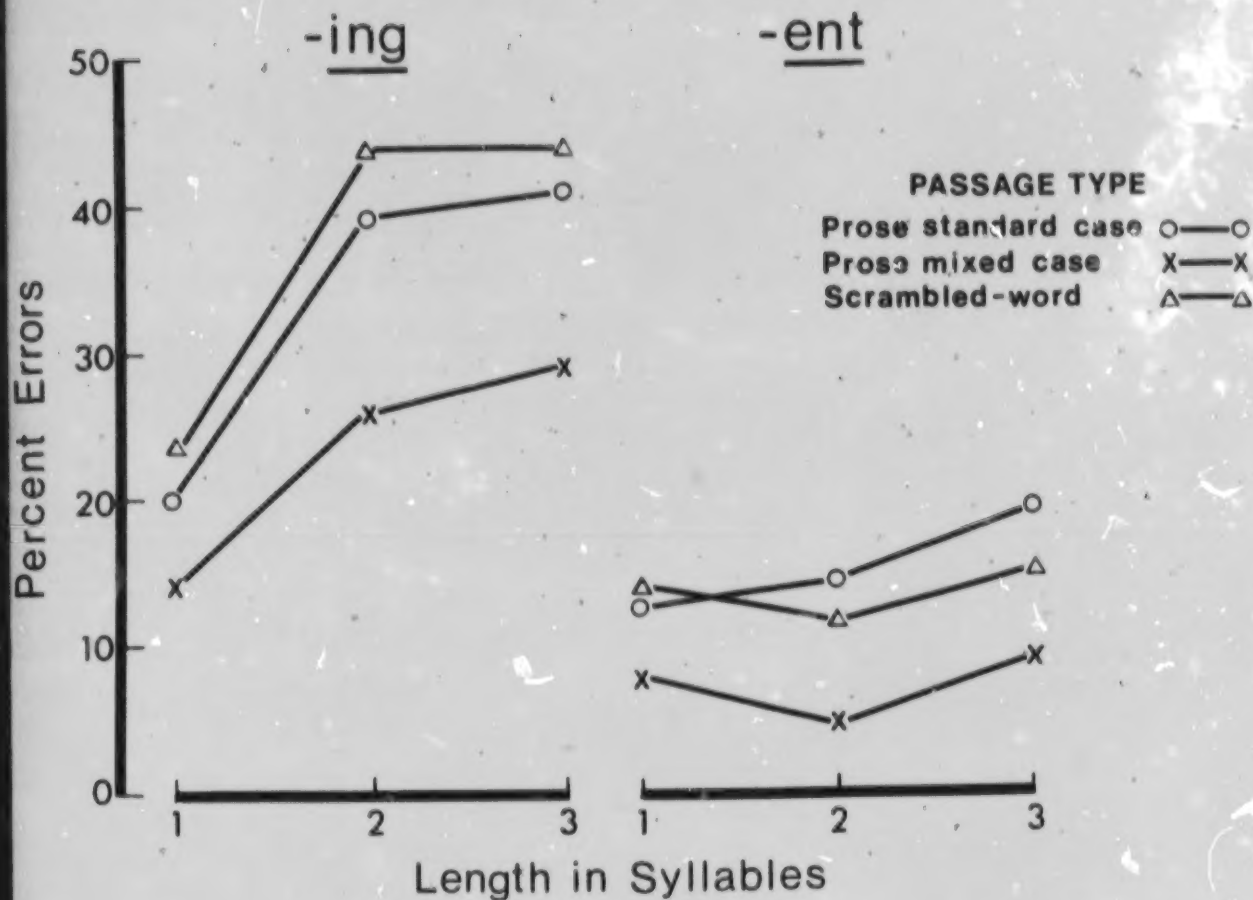


Figure 1. Error percentages for words ending in -ing or -ent in Experiment 1 as a function of the passage type and the length in syllables of the test words.

significant main effects of the type of word ending [$F(1,53) = 50.76$, $p < .01$], and the type of search passage [$F(2,106) = 14.77$, $p < .01$]. The subjects' performance was similar on prose standard case and scrambled-word passages and was uniformly more accurate on the prose mixed case passage, particularly for -ent endings. The interaction of word ending by passage type [$F(2,106) = 4.72$, $p < .05$] was significant.

A more detailed analysis of the data is presented in Figure 1, which shows error percentages for words ending in -ing or -ent at each syllabic length and with passage type as parameter. Error frequencies increased together with the length in syllables of content words ending in -ing, but no such effect was obtained for words ending in -ent. The main effect of the length in syllables [$F(2,106) = 28.22$, $p < .01$] and the interaction between the type of ending and syllabic length [$F(2,106) = 22.56$, $p < .01$] were both significant. The effect of syllabic length was greatest between mono- and disyllabic words, suggesting that the effects of syllabic length might reflect differences in the role of the test trigram. Whereas -ing (and -ent) trigrams in monosyllabic words necessarily formed part of the word stem, in di- and trisyllabic words, they mostly functioned as suffixes.

The incidence of errors on disyllabic test words ending in -ing was found to depend on whether the -ing trigram formed part of the word stem (e.g., during) or was a suffix morpheme (e.g., having). Data in Table 2 indicate that more letter detection errors were made when the -ing trigram was a suffix than when it was part of the word stem for the prose standard case [$t(53) = 2.54$, $p < .05$] and prose mixed case passages [$t(53) = 2.13$, $p < .05$]. This difference was in the same direction but was not significant for the scrambled-word passage [$t(53) = 0.11$, $p > .10$].

Table 2

Means (and Standard Errors) for Error Percentages in
Disyllabic -ing Content Words in Experiment 1
as a Function of Passage Type and Word Ending

Passage type	Word ending	
	stem	suffix
Prose standard case	32.6 (5.0)	45.7 (4.7)
Prose mixed case	21.3 (4.0)	30.0 (4.0)
Scrambled-word	43.7 (5.0)	44.3 (5.0)

The location of stress did not affect the incidence of errors on words ending in -ent. Instances of the target n in disyllabic words carrying stress on the first syllable (e.g., recent) were missed at the same rate as letters in words carrying stress on the second syllable (e.g., resent) for the prose

standard case [$t(53) = 0.98, p > .10$], prose mixed case [$t(53) = 1.52, p > .10$], and scrambled-word [$t(53) = -0.77, p > .10$] passages. The high incidence of errors on the -ing endings thus cannot be attributed to the fact that they never receive the primary word stress in English, since the incidence of errors on -ent endings, stressed or unstressed, was still appreciably lower than that on -ing endings.

Discussion

Substantially more letter detection errors were made on content words with -ing endings than with -ent endings. These data fail to support the notion that the high percentage of errors on -ing endings is solely due to the high orthographic frequency of the -ing sequence, since the trigram -ent is reported to have an even higher embedded frequency in English words than the trigram -ing (Underwood & Schulz, 1960). The data also fail to support the notion that the observed effects are solely due to the positional frequency of the target letter within the word, since the letter n always occurred in the penultimate location in both types of test words. Instead, the data suggest that the more important factor may be the positional frequency in the language of the entire trigram. Whereas the letter sequence -ent can occur in English at a number of locations within the word, the sequence -ing is found most often at word endings. The data are thus consistent with the hypothesis that the positional frequency of the letter sequence may be more critical than its absolute orthographic frequency.

Another important factor seems to be the linguistic function of the letter sequence: An analysis of errors made on disyllabic words showed that more errors were made when -ing endings were morpheme suffixes than when they formed part of the word stem. This finding also argues against a purely phonological interpretation of our results, since the n in the letter sequence -ing is pronounced in an atypical manner in both cases.

The subjects made fewer errors on the prose mixed case than on the prose standard case passage. Case mixing reduced but did not entirely eliminate the difference in errors between -ing and -ent endings and between suffix and word-stem groups. These data suggest that although purely visual manipulations influence error rates, visual factors alone cannot account for the observed pattern of results. Although error rates for -iNg and -InG were reduced relative to -ing, they were still above those for -ent.

There were no differences in error percentages on content words between the prose standard case and scrambled-word passages. This finding, which is consistent with our earlier reports (Healy, 1976; Drewnowski & Healy, 1977), suggests that phrases or word frames may be the largest units employed in fluent reading and that the presence of clauses or sentences has no further effect on the letter detection task. Since short-range syntactic/semantic constraints of text in the scrambled-word passage made it equivalent to the prose standard case passage in terms of the subjects' performance, we limited ourselves to scrambled-word passages in the subsequent experiments of this report.

EXPERIMENT 2

Data from Experiment 1 suggest that the positional frequency of the test trigram is critical in determining the percentage of letter detection errors. To test this hypothesis we studied variations in the location of the trigram within the word: Instances of n embedded in -ing or -ent trigrams now occurred in the middle or at the end of di- and trisyllabic content words. If the positional frequency of the trigram is the critical variable, then we would expect a large difference in error percentages with location for the more spatially predictable -ing trigrams and a smaller difference for -ent trigrams.

We also examined the possible effects of the syntactic/semantic structure of the search passage on the letter detection task. Because the context of the test words was not systematically varied in Experiment 1, the data did not rule out the possibility that the occurrence of -ing suffixes could be predicted on the basis of the words surrounding the test content words. It may be argued that the combination of surrounding context and test word stem allows the subject to anticipate and therefore skip over the word ending. We therefore presented the subjects with two passages of nonsense text differing in the appropriateness of the context surrounding the test content words (see Drewnowski & Healy, 1977, Experiment 2, for a similar manipulation using function words). Each passage was typed either in standard typecase or with alternating letters typed in upper and lower case. As in Experiment 1, this last manipulation was intended to impede the formation of reading units larger than letters using purely perceptual means.

Method

Subjects. Ninety-six male and female students at the University of Toronto served as volunteer subjects in a classroom setting. There were 53 subjects in the Standard Case condition and 43 different subjects in the Mixed Case condition. The two conditions were conducted in separate sessions.

Design and materials. Two 272-word passages of nonsense text were constructed. Each passage included 12 di- and 12 trisyllabic test words containing the trigram -ing, and 12 di- and 12 trisyllabic test words containing the trigram -ent. The location of the test trigram within the word was also systematically varied. In one half of the test words of each type and length, the -ing and -ent trigrams occurred at the end of the word, whereas in the remaining half of the test words, the -ing and -ent trigrams occurred in the middle of the word, either in the first or in the second syllable.

For di- and trisyllabic words ending in -ing, the trigram -ing was always an inflectional suffix (working, spelling, walking, hearing, falling, reading, performing, competing, amusing, exciting, amazing, visiting). In contrast, the trigram ending -ent either formed part of the word stem or occurred as a derivational suffix (recent, current, frequent, garment, absent, basement, competent, evident, vehement, excellent, resident, different). Selection of the test words was further constrained in such a way that the mean frequency of usage (Kućera & Francis, 1967) of test words containing -ing trigrams (28.0) was lower than that of words containing -ent trigrams (53.8).

The nature and order of filler words surrounding each test word were also systematically varied. Each test word was the central word of a three-word meaningful phrase (e.g., late visiting hours, the garment district). In addition, 24 filler phrases that were outwardly similar in construction to the test phrases (e.g., pure carrot juice, this chocolate cake) were randomly interspersed throughout the text. Each 272-word passage thus contained 48 three-word phrases containing test words, 24 three-word filler phrases, three instances each of the function words and and in, as well as 50 of the most frequent function words selected from the Kučera and Francis list with the restriction that no word include the letter n. The letter n thus always occurred in each passage in a function word in or and, or as part of an -ing or -ent string.

In the "local context passage" (cf. Drewnowski & Healy, 1977, Experiment 2), the words in each of the test phrases occurred in their correct order (late visiting hours), but the words in each of the filler phrases were reversed (juice carrot pure). Punctuation marks (periods) were inserted to make the text appear more like prose.

In the "no context passage," the words in each of the test phrases were reversed (hours visiting late), whereas the words in the filler phrases occurred in their correct order (pure carrot juice). This manipulation of the order of words within filler phrases was intended to bring both passages to roughly the same order of approximation to English. Punctuation and the location of the test words on the printed page remained the same for both passages. The context of the function words in and and within each passage was not systematically varied.

The two passages, local context and no context, were typed either in the standard fashion (Standard Case condition), or with alternating letters typed in upper and lower case (Mixed Case condition). There were two versions of each passage in the Mixed Case condition: Version A had all even letters in upper case and version B had all odd letters in upper case, so that across versions all instances of the target letter n occurred an equal number of times in lower and upper cases.

The two passages, presented in counterbalanced order and preceded by a page of instructions to subjects, were stapled together as a booklet. Approximately half the booklets in the Mixed Case condition contained version A of the two passages and the remaining booklets contained version B.

Procedure. Instructions to subjects and details of experimental procedure were the same as those described for Experiment 1.

Results

The data are summarized in Table 3, which includes means and standard errors of the means for error percentages for each type of test word and under each set of experimental conditions.

Table 3

Means (and Standard Errors) for Error Percentages in Experiment 2
as a Function of Condition, Passage Type, and Word Type

Condition	Word type			
	Function		Content	
	<u>in</u>	<u>and</u>	<u>ing</u>	<u>ent</u>
Standard Case				
Local context	64.7 (3.0)	59.0 (5.0)	32.7 (3.4)	15.9 (2.0)
No context	69.7 (4.3)	72.3 (3.7)	34.3 (3.4)	14.4 (2.0)
Mixed Case				
Local context	54.0 (4.7)	26.3 (2.7)	21.7 (3.0)	9.1 (1.4)
No context	56.3 (5.3)	35.0 (4.0)	19.9 (2.8)	6.4 (1.6)

Function words. More errors were made on function words in and and than on content words, in agreement with previous results (e.g., Drewnowski & Healy, 1977). Both function words were more likely to be detected when in the Mixed Case than in the Standard Case condition, and this improvement was much greater for the word and than for the word in. The main effect of test word type (in or and) was significant in the Mixed Case [$F(1,42) = 48.35, p < .01$], but not in the Standard Case [$F(1,52) < 1$] condition. A more detailed treatment of this phenomenon is presented in Experiment 4.

Content words. More errors were made on test words containing -ing trigrams than on words containing -ent trigrams both in the Standard Case [$F(1,52) = 60.05, p < .01$] and in the Mixed Case [$F(1,42) = 48.71, p < .01$] conditions. In contrast, and contrary to our hypothesis about the possible effects of syntactic/semantic contexts on the detection tasks, the effects of passage context [Standard Case: $F(1,52) < 1$; Mixed Case: $F(1,42) = 2.52, p > .10$] were not significant. Furthermore, the effect of the length of the test words in syllables [Standard Case: $F(1,52) = 1.97, p > .10$; Mixed Case: $F(1,42) = 3.84, p > .05$] was not significant. More errors were made when the two trigrams occurred at the end rather than in the middle of the test words. The main effect of target location was significant for both the Standard Case [$F(1,52) = 97.13, p < .01$] and the Mixed Case [$F(1,42) = 42.55, p < .01$] conditions.

The data, averaged over passage context and test word length in syllables, are shown in Table 4 to illustrate the nature of the interaction of trigram type by location, which was significant both for the Standard Case [$F(1,52) = 61.70, p < .01$] and for the Mixed Case [$F(1,42) = 28.18, p < .01$] conditions. Whereas the percentages of letter detection errors on -ing and on -ent trigrams were approximately equal when the trigrams occurred in the

middle of the test word, significantly more errors were made on -ing trigrams than on -ent trigrams when the trigrams occurred as word endings. The same interaction held for both the Standard Case and Mixed Case conditions, although overall error percentages in the Mixed Case condition were greatly reduced.

Table 4

Means (and Standard Errors) for Error Percentages in Experiment 2
as a Function of Condition, Test Trigram, and Target Location

Condition	Test trigram			
	<u>ing</u>		<u>ent</u>	
	Middle	End	Middle	End
Standard Case	15.9 (2.5)	50.7 (5.0)	10.4 (1.9)	20.0 (2.6)
Mixed Case	11.4 (2.5)	30.2 (3.9)	6.1 (1.6)	9.2 (1.9)

Discussion

The present finding that more detection errors were made in later rather than in earlier target locations within the word replicates the location effects observed by Corcoran (1966). This effect was small when the target letter n was part of the letter sequence -ent and much greater when it was part of the sequence -ing. This finding of a strong interaction between the type of the test trigram and its location within the word indicates that the orthographic, or linguistic, context of the target letter must also be a contributing factor. Because the location of the target letter within the word was controlled across both types of test trigrams, we are able to rule out the possibility that single-letter positional frequency is the basis for the subjects' responses (Katz, 1977; Mason, 1975). Rather, our results point to another kind of frequency mediation--based not so much on orthographic frequency of letter sequences, but instead on their positional frequency. In other words, our results suggest that frequent letter sequences may act as reading units only if they occur in their most common locations within the word.

Although the within-word context of the target letter was found to be of importance, the present data show no trace of between-word context effects as determined by the syntactic/semantic structure of the search passage. These data qualify the notion that the combination of surrounding context and test word stem allows the subject to pass over the word ending. Because manipulations of passage context in our previous studies influenced the pattern of

errors on the function words the and and (Drewnowski & Healy, 1977), we were led to postulate that very frequent function words are processed in terms of reading units larger than the word itself. The present data do not permit a similar conclusion for the less frequent content words, since equivalent results were obtained for the local context and no context passages. Instead, it may be that in the case of multisyllabic test words, the reading unit is larger than the letter but smaller than the word. Consequently, we need to examine more closely the possibility that the -ing suffix morpheme itself serves as a processing unit in reading multisyllabic words.

EXPERIMENT 3

In Experiment 3 we asked one group of subjects to search for the letter n and another group of subjects to search for the entire -ing trigram embedded in test words. We wished to examine whether word endings per se are harder to identify than other parts of words, or whether component letters in word endings are "hidden" when the endings are read in terms of units larger than the letter. If word endings are harder to identify, then we should expect more errors on word endings than on mid-word locations for both types of search targets. In contrast, if the suffix morpheme serves as a reading unit, then we should expect more errors on word endings than on mid-word locations only for the search target n, not for the search target -ing. Only in the former case would the target be smaller than the reading unit employed; in the latter case, the reading unit and the to-be-detected target would be identical.

Method

Subjects. One hundred and sixty students at the University of Toronto served as volunteer subjects in the present experiment, which was conducted in the classroom. There were 80 subjects in the Letter Detection condition and 80 subjects in the Trigram Detection condition.

Design and materials. One 272-word passage of nonsense text, derived from the local context passage of Experiment 2, was used. The three instances each of function words in and and as well as the 24 -ent test words had been removed and replaced by words of equal length and, where possible, of equivalent frequency that did not contain the letter n. The punctuation and the location of the test words on the page remained the same as in Experiment 2, though the overall target density (24 targets per 272 words) was necessarily reduced. All the remaining instances of the letter n were part of the -ing trigram.

The present passage was then typed either in the standard fashion (standard case passage) or with alternating letters typed in upper and lower case (mixed case passage) in the same manner as the passages in Experiment 2.

Procedure. Each subject received both the standard and mixed case passages presented in a counterbalanced order. Details of experimental procedure otherwise resembled those of Experiment 2, except that half the subjects were instructed to search for the letter n and half the subjects were instructed to search for the trigram -ing.

Results

The data are summarized in Table 5, which shows error percentages for each detection task at each target location within the word and with passage type as parameter. These data have been pooled over the syllabic length of the test words.

Table 5

Means (and Standard Errors) for Error Percentages in Experiment 3
as a Function of Search Task, Passage Type, and Target Location

Passage type	Search task <u>n</u>		<u>ing</u>	
	Middle	End	Middle	End
Standard case	16.8 (2.0)	53.5 (3.8)	12.9 (2.0)	9.8 (1.1)
Mixed case	14.5 (1.8)	41.5 (3.4)	11.5 (2.2)	7.6 (1.1)

The subjects made significantly more errors on the letter detection than on the trigram detection task [$F(1,158) = 69.19, p < .01$]. Subjects searching for the letter n made significantly more errors at end locations than at middle locations (in agreement with the data of Experiment 2), whereas subjects searching for -ing trigrams made somewhat fewer errors at end locations. Both the main effect of location [$F(1,158) = 73.53, p < .01$] and the critical interaction between search task and location [$F(1,158) = 114.13, p < .01$] were significant.

The subjects made more errors on standard case than on mixed case passages [$F(1,158) = 14.88, p < .01$]. This difference was greater in the letter than in the trigram detection task, particularly for end targets. The interaction between typeface and location [$F(1,158) = 9.39, p < .01$] as well as the second order interaction between the search task, typeface, and location [$F(1,158) = 6.97, p < .01$] were significant.

Discussion

These data show that most letter detection errors were made on normally typed -ing endings by subjects searching for the target letter n. In contrast, subjects instructed to search for the entire -ing trigram were much more accurate in their performance and were influenced neither by target location within the word nor by the passage typeface. This pattern of results differs from that obtained previously with common function words (Drewnowski &

Healy, 1977), in which subjects made a disproportionate number of errors on the word and whether they were instructed to search for the letter n or the entire word and.

These results point to an important difference in processing content and function words. We have interpreted our previous data on function words by proposing that detection errors resulted whenever the size of the target was smaller than the size of the postulated reading unit. Thus a disproportionate number of letter detection errors on the word and implied that it was processed in terms of units larger than the letter. Similarly, the high percentage of detection errors when the entire word and was the target implicated the use of reading units larger than the word itself (Drewnowski & Healy, 1977). Following this reasoning, we might expect a sharp drop in the percentage of detection errors whenever the size of the target is equal to or is larger than the size of the postulated reading unit.

We therefore interpret the present data by proposing that -ing endings can serve as single units in the course of fluent reading. Most errors were made when the size of the postulated reading unit exceeded that of the to-be-detected target. This occurred for letter targets when the -ing trigram was in its most frequent location within the word. Fewer errors were made on mixed case passages in which the formation of -ing reading units was to some extent impeded. Few errors were also made for the target -ing, which either corresponded in size to our postulated reading unit, as it did in standard case passages, or exceeded the size of the available reading unit, as it did in mixed case passages.

An additional factor that may play a role when the subjects search for -ing targets is that subjects may scan word endings more thoroughly, since -ing occurs predominantly at the end of words in English. Whereas Experiment 2 suggested that subjects are, if anything, less likely to process the final -ing letter by letter when searching for ns, they might look more carefully at the word endings when searching for -ing, since -ing is a unit expected to occur at the end of words. This hypothesis implies that the subjects are sensitive to the position of letter sequences within words, in the manner postulated for single letters by Mason (1975) and explored for letter trigrams in Experiments 1 and 2.

EXPERIMENT 4

We have now shown that subjects make many detection errors on the letter n when it occurs as part of an -ing ending. Because such a pattern of errors is not found for -ent endings, we proposed that the critical factor determining the subjects' performance is the positional frequency of the letter sequence. However, there may be an alternative linguistic explanation for this result. First, the test words in Experiments 1 and 2 were such that -ing endings generally served as suffix morphemes, whereas this was not always the case with -ent endings. Second, -ing suffixes were always inflectional, whereas -ent suffixes were derivational (Jarvella & Snodgrass, 1974). Thus, -ing suffix test trigrams occurred in inflected forms of verbs, whereas -ent suffix test trigrams occurred in derived adjectives and nouns. Third, the inflectional and derivational suffixes were paired with different word stems.

We now wish to examine the importance of each of these factors. We therefore consider several suffix morphemes other than -ing to determine whether they also can function as perceptual units in reading. Specifically, we compare inflectional and derivational suffixes attached to the same word stems.

Work in linguistics has suggested that bound morphemes vary in their degree of attachment to the word stem and that derivational suffixes are more tightly bound than inflectional suffixes (Kean, 1977). It seems reasonable to postulate that inflectional suffix morphemes are more likely to function as single linguistic entities and hence are more likely to be unitized during reading. We therefore compare error percentages on inflectional suffixes, such as -ing and -en, to those on derivational suffixes, such as -ment and -ion. In each case, test words ending in -ing contained the same word stems as test words with each of the other three endings.

Method

Subjects. One hundred and sixteen male and female Yale undergraduates participated as volunteer subjects in a group experiment conducted in the classroom.

Design and materials. Two 100-word scrambled-word passages were constructed. The first passage, which will hereafter be referred to as the "standard case passage," included six words ending in -ment (e.g., development) and the corresponding six words (the six words with the same stem) ending in -ing (e.g., developing), six words ending in -en (e.g., taken) and the corresponding six words ending in -ing (e.g., taking), six words ending in -ion (e.g., suggestion) and the corresponding six words ending in -ing (e.g., suggesting). Each of the two words with the same stem differed only in the suffix. We therefore have three word-stem groups: -ment, -en, and -ion. Note that two of the endings employed (-ment and -ion) were derivational endings and two (-ing and -en) were inflectional endings. Also note that for two of the endings (-en and -ion), the target letter n was the final letter and for the other two endings (-ing and -ment), the target letter n was the penultimate letter. In addition to these 36 words containing target letters, each passage contained three instances of the word and and three instances of the word in.

For each of the 42 test words, a word identical in length and comparable in frequency, according to Kučera and Francis (1967), but without the letter n was selected for inclusion in the passage as a filler. The mean frequency of the six test words containing a suffix of a given type was matched exactly to the mean frequency of the six analogous filler words. Similarly, there were three instances of the filler function word was and three of the filler function word of, which were matched to the instances of the test function words and and in, respectively. These constraints were instituted to insure that subjects could not determine whether a word contained a target letter on the basis of its length, its frequency in the language, or its frequency in the passage. An additional 16 function words were selected as fillers to make the passage closer to natural prose. The order of the words in the passage was pseudorandom with the constraint that the first two and last two words were selected from the 16 function-word fillers and each remaining block of 16

successive words included one of each of the six types of test word with suffix and the matched filler words, one instance of either and or in and the matched filler word, and two of the additional function-word fillers. The order of words within a block of sixteen was random.

The second passage, hereafter referred to as the "mixed case passage," was identical to the standard case passage except that every other letter was typed in capitals. There were two versions of this passage. In version A even letters were capitalized, and in version B odd letters were capitalized.

Each passage was typed on a separate sheet of paper, and each subject was shown a copy of the standard case passage and a copy of one of the two versions of the mixed case passage, presented in a counterbalanced order and preceded by a sheet of instructions. The two passages and the instruction sheet were stapled together. Half of the subjects were shown version A of the mixed case passage and half were shown version B.

Procedure. The instructions and procedure were essentially the same as in the preceding experiments.

Results

The results are summarized in Table 6, which shows the mean and the standard error of the mean error percentages as a function of passage type and test word category.

Table 6

Means (and Standard Errors)
for Error Percentages in Experiment 4
as a Function of Passage Type and Word Category

Passage type	Word-Stem Group							
	Function		<u>ment</u>		<u>en</u>		<u>ion</u>	
	<u>and</u>	<u>in</u>	<u>ing</u>	<u>ment</u>	<u>ing</u>	<u>en</u>	<u>ing</u>	<u>ion</u>
Standard case	41.7 (3.7)	51.0 (3.3)	27.6 (2.5)	13.1 (2.0)	23.0 (2.5)	7.0 (1.1)	21.6 (2.5)	6.2 (1.2)
Mixed case	14.0 (2.7)	38.7 (3.3)	9.3 (1.6)	4.9 (1.2)	9.2 (1.8)	3.9 (1.0)	8.0 (1.6)	3.7 (1.1)

Function words. As in the previous experiments, subjects made more errors on the common function words and and in than on the other test words. Also, as noted previously, the percentage of errors on the function words was greater in the standard case passage than in the mixed case passage [$F(1,115)$]

= 38.02, $p < .01$). The difference between the two passages was much larger for the word and than for the word in [$F(1,115) = 13.91$, $p < .01$]. A more detailed analysis of the data provides an explanation for this difference between and and in. In version A of the mixed case passage, each of the three instances of the word in was typed as In, which is a regular version of the word, whereas in version B, each of the three instances of the word in was typed as iN, which is not a regular version of the word. The mean error percentage on the word in was much larger in version A of the mixed case passage (mean = 52.3) than in version B (mean = 24.7) [$t(114) = 4.55$, $p < .01$]. This analysis suggests that only when the word in is typed in an irregular fashion will the percentage of letter-detection errors on it be depressed, which is consistent with the hypothesis that subjects make many letter-detection errors on the word in only when they are able to read it in units larger than the letter.

Content words. For each word-stem group, test words ending in -ing yielded more errors than the corresponding words ending in the alternative suffixes [$F(1,115) = 61.05$, $p < .01$]. There was also an effect of word-stem group: More errors were made on the words in the -ment group than in the other two groups [$F(2,230) = 12.59$, $p < .01$], and the group factor did not interact with the factor of suffix type (-ing vs. the alternative suffixes) [$F(2,230) < 1$]. This result suggests that the different error frequencies among the three word-stem groups were not due to differences in word endings but rather to differences in word stems.

As observed in the earlier experiments, subjects made more errors on the standard case passage than on the mixed case passage [$F(1,115) = 49.2$, $p < .01$]. The difference between passages was greater for the words ending in -ing than for the corresponding words with other endings, and was somewhat greater for the -ment group than for the other two groups. The interaction between passage type and suffix type (-ing vs. the alternative suffixes) was significant [$F(1,115) = 43.16$, $p < .01$], as was the interaction between passage type and word stem group [$F(2,230) = 8.30$, $p < .01$]. Hence, the difference between passages was greatest for those conditions with most errors.

Discussion

Subjects made more errors on words with -ing endings than on words with any other ending. Since the positional frequency of -ing at the end of a word is considerably greater than that of any of the other letter sequences used, the present results provide further support for unitization of letter groups on the basis of their positional frequency. Spatial predictability on its own cannot account for the present pattern of results because -ion, like -ing, occurs almost exclusively at the end of words (Mayzner, Tresselt, & Wolin, 1965). However, the letter sequence -ing is more frequent than the letter sequence -ion (Underwood & Schulz, 1960; -ing: 1673, -ion: 1370). Hence, orthographic frequency would seem to be a prerequisite, though not a sufficient condition, for the unitization of letter groups.

Although the results of Experiments 1 and 2 show that more errors were made on word suffixes than on word stems, the present task does not discriminate between different classes of suffixes. Specifically, we found no

consistent differences in error percentages between inflectional and derivational suffixes. Since the words with the inflectional ending -en yielded a similar number of errors as the words with the derivational ending -ion and fewer errors than the words with the derivational ending -ment, the preponderance of errors on the words ending in -ing cannot be attributed solely to the fact that -ing is an inflectional, as opposed to a derivational, ending. Furthermore, since the same word stems were used with -ing endings as with the other endings, the differences found between errors on the words ending in -ing and errors on the words with other endings cannot be attributed to differences in the word stems. Finally, since not all suffixes yielded many detection errors, these data argue against the existence of a suffix as the sole explanation of the observed effects; other factors must be involved as well.

The fact that the errors on the words ending in -ing were particularly disrupted by typing the words with every other letter in capitals further supports the hypothesis that the errors were caused by the subjects' reading the suffix -ing as a perceptual unit.

SUMMARY AND CONCLUSIONS

Subjects searching for the target letter n made many letter detection errors on content words ending with the suffix morpheme -ing. In this regard the suffix morpheme -ing resembled the frequent function words in and and, although the percentage of errors was smaller for -ing than for either of the two function words. We focused on three factors that may contribute to the missing letter effect: (1) the orthographic frequency of the letter sequence -ing, (2) its spatial predictability within the word, and (3) its linguistic function as a word boundary morpheme. Our data suggest that although each of these factors may be important for the missing letter effect, no factor alone is sufficient. Whereas whole word frequency has been shown to be critical to the missing letter effect (Healy, 1976), the orthographic frequency of the test trigram cannot be the sole contributing factor, since considerably fewer errors were made on words ending with the letter sequence -ent than with the sequence -ing, which is reported to be less frequent (Experiments 1 and 2). Furthermore, fewer errors were made on -ing when it occurred in the middle of the word than when it was the word ending (Experiment 2). This result agrees with our earlier studies in which considerably fewer errors were made on letter sequences the and and when they were embedded in other words (e.g., thesis, handle) than when they occurred as separate function words. The high orthographic frequency of the letter sequence in which the target is embedded may be a necessary condition for the appearance of the missing letter effect but is in itself not sufficient.

Similarly, although the results of Experiment 1 suggest that the spatial predictability of the target trigram may be important, it cannot be the sole contributing factor to the missing letter effect for two reasons. First, fewer errors were made on words ending in -ion than on words ending in -ing, despite the fact that -ion, like -ing, occurs almost exclusively at the end of English words (Experiment 4). Second, fewer errors were made on the -ing ending itself when it occurred as part of a word stem than when it served as a suffix morpheme (Experiment 1). However, the fact that the letter sequence

was a suffix morpheme could not be the only critical factor, since considerably fewer errors were made on the suffix morphemes -ment, -en, and -ion than on -ing (Experiment 4). We tentatively conclude that it is the combination of orthographic frequency, spatial predictability, and linguistic role as a word boundary morpheme that may be necessary for full expression of the missing letter effect.

In addition, we have identified a number of other factors that may affect the rate of detection errors in this task but cannot account for the specific pattern of errors in the present study: (1) the single letter spatial predictability, (2) the nature of the attached word stem, (3) the frequency of the test word, (4) the syntactic and semantic context of the test word, and (5) the class of the suffix morpheme (inflectional vs. derivational). First, subjects in Experiment 1 made many more errors on words ending in -ing than in -ent, despite the fact that n was the penultimate target letter in both cases. Second, subjects in Experiment 4 made more errors on words ending in -ing than on words ending in -ment, -en, or -ion, despite the fact that the same word stem was used for -ing as for the other endings. Third, subjects in Experiment 2 made more errors on words containing -ing than -ent, despite the fact that the mean frequency of words containing -ing was lower than that of words containing -ent. Fourth, manipulations of test word context in Experiment 2 did not affect the number of letter detection errors. Finally, manipulating the class of the suffix morphemes in Experiment 4 did not yield consistent differences between inflectional and derivational suffixes.

We have thus isolated some of the conditions responsible for the large number of letter-detection errors on the sequence -ing. We propose that these errors occurred because under these conditions -ing tends to be unitized during reading and its constituent letters are thereby "concealed." The most direct support for the hypothesis that the letter sequence -ing may serve as a single unit during reading is provided by the data of Experiment 3. Subjects searching for the letter sequence -ing made fewer detection errors than subjects searching for the letter n in the same passage. Such a pattern of results would be expected to occur if the n target but not the -ing target was smaller than the reading unit employed. Furthermore, fewer errors were made by subjects searching for the letter n in passages of mixed typecase than in normally typed passages, particularly for words with -ing endings. This finding would be expected to occur if the formation of the reading units is disturbed by the use of mixed typecase. The fact that the differences in error rates between different word endings are reduced but not eliminated by the use of mixed typecase suggests that the postulated reading units may be response units, rather than perceptual, or purely visual, ones.

The present data may be limited in their generality, because few if any suffix morphemes other than -ing may meet the constraints we have specified above for the occurrence of the missing letter effect. However, these data taken in conjunction with the results previously obtained for function words (Drewnowski & Healy, 1977) and for common nouns (Healy, 1976, 1980) provide a coherent and general framework in which to understand the nature of the reading process. In particular, this series of studies provides evidence that subjects employ units of varying sizes in reading passages of text. Earlier we showed that common words are more likely to be read in units larger than the letter than are rare words and that frequent function words in appropriate

linguistic context may be processed in units that include more than one word. We have now shown that frequent morpheme suffixes may also be read as single units. In addition, less frequent word stems may be read in terms of letter units. We therefore propose that in reading a given passage, the subject can employ units of different sizes, not only within the passage, but also within a single word.

REFERENCES

- Corcoran, D. W. J. An acoustic factor in letter cancellation. Nature, 1966, 210, 658.
- Drewnowski, A. Detection errors on the word the: Evidence for the acquisition of reading levels. Memory & Cognition, 1978, 6, 403-409.
- Drewnowski, A., & Healy, A. F. Detection errors on the and and: Evidence for reading units larger than the word. Memory & Cognition, 1977, 5, 636-647.
- Gibson, E. J. Learning to read. Science, 1965, 148, 1066-1072.
- Healy, A. F. Detection errors on the word the: Evidence for reading units larger than letters. Journal of Experimental Psychology: Human Perception and Performance, 1976, 2, 235-242.
- Healy, A. F. Proofreading errors on the word the: New evidence on reading units. Journal of Experimental Psychology: Human Perception and Performance, 1980, 6, 45-57.
- Jarvella, R. J., & Snodgrass, J. G. Seeing ring in rang and retain in retention: On recognizing stem morphemes in printed words. Journal of Verbal Learning and Verbal Behavior, 1974, 13, 590-598.
- Johnson, N. F. On the function of letters in word identification: Some data and a preliminary model. Journal of Verbal Learning and Verbal Behavior, 1975, 14, 17-29.
- Kean, M.-L. The linguistic interpretation of aphasic syndromes: Agrammatism in Broca's aphasia, an example. Cognition, 1977, 5, 9-46.
- Katz, L. Reading ability and single-letter orthographic redundancy. Journal of Educational Psychology, 1977, 69, 653-659.
- Kučera, H., & Francis, W. N. Computational analysis of present-day American English. Providence, R.I.: Brown University Press, 1967.
- Locke, J. L. Phonemic effects in the silent reading of hearing and deaf children. Cognition, 1978, 6, 175-187.
- Mason, M. Reading ability and the letter search time: Effects of orthographic structure defined by single-letter positional frequency. Journal of Experimental Psychology: General, 1975, 2, 146-166.
- Mayzner, M. S., Tresselt, M. E., & Wolin, B. R. Tables of trigram frequency counts for various word-length and letter-position combinations. Psychonomic Monograph Supplements, 1965, 1, 33-78.
- Thorndike, E. L., & Lorge, I. The teacher's word book of 30,000 words. New York: Columbia University, Teachers College, Bureau of Publications, 1944.
- Underwood, B. J., & Schulz, R. W. Meaningfulness and verbal learning. New York: Lippincott, 1960.
- Wheeler, D. D. Processes in word recognition. Cognitive Psychology, 1970, 1, 59-85.

FOOTNOTES

¹The word during may seem questionable as an example of a word stem; however, this should only minimize any differences between the word stem and the morpheme suffix groups of words.

²The present scrambled-word passage is comparable to the scrambled-word passages of Healy (1976) and Drewnowski and Healy (1977, Experiment 1) and not to the scrambled-word passage of Drewnowski (1978), in which every instance of the test word the occurred in an inappropriate context.

152

SPEAKING RATE, SYLLABLE STRESS, AND VOWEL IDENTITY*

Brad Rakerd+, Robert R. Verbrugge+, and Donald P. Shankweiler+

Abstract. Increases in speaking rate and decreases in syllable stress have been thought to be accomplished with a common production strategy, and correspondingly, to produce comparable variations in vowel acoustics. If this is the case, it would be expected that a listener makes similar use of rate and stress information in vowel perception. In this study, the perceptual influences of rate and stress information were assessed for the perception of vowels in /pVp/ syllables. For each of nine vowels (i, I, E, æ, a, A, ɔ, u, u) a male talker recorded the carrier sentence "I think it's the yellow /pVp/'s call" at slow and fast rates, while placing primary stress on the target /pVp/ or upon the preceding word "yellow." Synthetic vocalic regions of /pVp/ syllables were then generated on an OVE-III synthesizer, pairing the formant frequency values of natural slow-stressed syllables with the pitch, amplitude, and duration of the corresponding syllables produced in slow-stressed, fast-stressed, and slow-destressed carriers. These synthetic syllables were then edited into the natural carriers and presented to listeners in a vowel identification task. The pattern of identification errors differed for the three types of sentence carriers: short-long vowel pairs were confused infrequently and with about equal probability for the original slow-stressed carrier sentences, while the fast-stressed carriers were characterized by a predominance of long vowel-short vowel confusions and the slow-destressed carriers by short vowel-long vowel confusions. This difference is taken as evidence that rate and stress transformations are not perceptually equivalent, perhaps reflecting differences in how these factors are regulated in production.

Increases in speaking rate and decreases in syllable stress have often been characterized as having comparable consequences for vowel acoustics. Lindblom (1963) observed, for example, that the vowels in fast speech and destressed speech are shorter and reduced in vowel color relative to the vowels in slow-stressed speech. He therefore proposed that a single production strategy is employed to effect rate and stress changes. Invariant motor

*This paper was presented at the 99th meeting of the Acoustical Society of America, Atlanta, Georgia, 21-25 April, 1980.

+Also University of Connecticut, Storrs.

Acknowledgment. This research was supported by NICHD grant HD-01994 and BRS grant RR-05596 to Haskins Laboratories.

commands are said to specify the articulatory targets that correspond to individual vowels, with rate and stress contrasts being defined by the time intervals between successive commands. In the case of rapid or destressed speech, these intervals would be too brief to allow the articulators to reach their targets, and this would result in shortened segment duration and reduced vowel color.

Acoustic measurements have not always shown an invariant pairing of shortened duration and spectral vowel reduction, however. We (Verbrugge & Shankweiler, 1977) found that the vowels in destressed /pVp/ syllables, spoken in the carrier sentence "I think it's the yellow /pVp/'s call," were not appreciably shorter than their stressed counterparts, despite the fact that the two sets of vowels displayed marked spectral differences. Other investigators (Harris, 1978; Gay, 1978) have found complementarily that an increase in speaking rate may result in reduced duration with little change in center formant frequency values. These data suggest that rather than being regulated with a common production strategy, rate and stress are independently regulated, and they may differentially affect vowel acoustics; correspondingly, they may differentially affect vowel perception.

In a previous study (Verbrugge & Shankweiler, 1977), we conducted a perceptual test of listeners' use of the information for rate and stress in their identification of vowels. An adult male talker produced nine vowels in /pVp/ syllables in four sentence contexts: the target syllables were either stressed or destressed in sentences that were spoken at fast or slow rates. Listeners were then asked to identify the vowels when syllables were presented in their original sentence context or in a different context.

Identification errors increased significantly when fast syllables were presented in slow sentences. Somewhat surprisingly, a complementary influence of fast sentence context on slow syllables was not observed; nor was an effect of stress context observed for stress interchanged syllables, despite marked spectral differences in the syllables produced at different levels of stress. These negative findings may reflect a limitation of the experimental design, however. The target syllables themselves contained prosodic information that reflected their original sentence context. We suspected that this information was sufficient to specify their original stress level and, in the case of the slow syllables, to specify their original rate of articulation, even when these syllables appeared in an inappropriate sentence context. To overcome this limitation, the vocalic regions of /pVp/ syllables were produced synthetically in the current study, so that spectral information and prosodic information could be manipulated independently. In this way, target syllables could be produced with a prosody that matched the carrier sentence, but with a formant structure appropriate to a different context.

Three sets of stimuli were produced. In each set, the formant values for synthesis were those of natural stressed syllables spoken in a carrier sentence at a slow rate, with the formant values of the vowels /i, I, E, A, a, A, J, v, u/ being measured at the pitch pulse of maximum amplitude. Measurements for the first and second formants are displayed in Figure 1 (the points connected by the solid line), and it can be seen that these formants are representative for an adult male talker (cf. Peterson & Barney, 1952).

One stimulus set synthetically paired the slow-stressed formants with the prosody of their original slow-stressed carrier sentences, while a second set paired the same formants with the prosody of fast-stressed carriers. The resulting synthetic syllables were then edited into their corresponding carrier sentences, and the perceptual consequences of increased speaking rate were assessed by comparing responses to these slow-stressed and fast-stressed sentences. In the third stimulus set, the slow-stressed formants were paired with the prosody of slow-destressed carriers. A comparison of responses to the slow-destressed and slow-stressed sentences indexed the perceptual effect of destressing.

Subjects were asked to identify the vowels in the target /pVp/ syllables in one of two listening tests: they heard either a randomized list of slow-stressed and fast-stressed sentences, or a randomized list of slow-stressed and slow-destressed sentences. Responses were indicated by circling one of nine "p--p" words written in English orthography.

The parameters used in syllable synthesis are illustrated in Figure 3. As mentioned, the formant values were those of naturally produced slow-stressed syllables, with formants one, two, and three being iterated for the duration of synthesis. The duration was specified by the duration of the vocalic region of the natural syllable produced in the original carrier. Fundamental frequency was measured at the points of onset and offset of voicing and linearly interpolated between these points. Amplitude was also measured at onset and offset of voicing, and at the pitch pulse of maximum amplitude. The amplitude contour was interpolated from onset to the maximum and from the maximum to offset. Synthetic segments were generated on an OVE-III synthesizer and edited into carriers immediately after the natural /p/ release and before final /p/ closure. This pairing of natural and synthetic speech was effective; no subject detected a change of source and all subjects reported that the test sentences sounded natural.

To summarize, the target syllables in the three stimulus sets contained identical spectral information, but the pitch, amplitude, and duration of the syllables varied to reflect changes in rate and stress. This design permitted a direct comparison of the influences of rate and stress information on a perceiver's sensitivity to the spectral information for vowels. If, for example, fast speech and destressed speech have comparable spectral consequences in production, then one would expect to find similar perceptual responses to the fast-stressed and slow-destressed sentence carriers.

Responses were analyzed both in terms of the absolute accuracy of vowel identification and in terms of the pattern of confusion errors. With regard to the overall accuracy of identification, responses to the fast-stressed and slow-destressed carriers were similar, in that both groups showed an overall increase in identification errors relative to the original slow-stressed carriers. (Identification errors averaged 13% for the fast-stressed and 9% for the slow-destressed carriers, compared to 6% for the original slow-stressed carriers.)

Responses to the fast-stressed and slow-destressed carriers differed markedly in terms of the overall pattern of confusion errors, however. Confusion errors for the vowels /ɛ/ and /æ/ are displayed in Figure 4. The

open bars represent confusions of /ɛ/ with /æ/ and the hatched bars represent confusions of /æ/ with /ɛ/. Both types of errors are shown for the original slow-stressed carrier sentences, for the fast-stressed carriers, and for the slow-destressed carriers. It can be seen that when the slow-stressed formants appeared in their original slow-stressed carriers, /ɛ/ and /æ/ were never confused. However, when the same formants appeared in fast-stressed carriers, there were confusions of both types, but particularly, confusions of /æ/ with /ɛ/. Finally, when the slow-stressed formants appeared in slow-destressed carriers, there were a number of /ɛ - æ/ confusions but no /æ - ɛ/ confusions. The pattern of confusion errors is clearly different for the fast-stressed and slow-destressed carriers.

It should be noted that the predominance of /æ - ɛ/ confusions in the fast-stressed carriers cannot be accounted for simply in terms of vowel duration. While it is the case that the fast-stressed syllables were shorter than the slow-stressed syllables, it is also the case that the fast syllables were presented in the context of rapidly articulated carriers. In this context, /æ/ was still relatively longer than /ɛ/. Presumably, then, the fast carriers provided sufficient information for accurately discriminating /ɛ/ and /æ/ on the basis of relative duration. Errors in the fast-stressed carriers are more likely due to alterations in perceptual sensitivity to spectral variables that distinguish /ɛ/ from /æ/.

Confusion errors for the vowels /ʌ/ and /ɑ/ are shown in Figure 5, with the open bars representing /ʌ - ɑ/ confusions and the hatched bars representing /ɑ - ʌ/ confusions. The pattern of errors is similar to that for the vowels /ɛ/ and /æ/. In the original slow-stressed carriers, /ʌ/ and /ɑ/ were confused infrequently and with equal probability. In the fast-stressed carriers, the frequency of confusions increased, and particularly, the frequency of /ɑ - ʌ/ confusions increased. Finally, in the slow-destressed carriers, while confusion errors dropped off overall, there were more /ʌ - ɑ/ than /ɑ - ʌ/ confusions.

It is evident that responses to the fast-stressed and slow-destressed carriers differ in terms of the direction of confusion errors, the fast-stressed carriers being characterized by a predominance of long vowel-short vowel confusions and the slow-destressed carriers by a predominance of short vowel-long vowel confusions. When responses are pooled over the short vowels /ɛ/ and /ʌ/, and the long vowels /æ/ and /ɑ/, an analysis of variance shows a significant interaction of carrier sentence type and the direction of confusion errors, $F(1,22)=7.72$, $p<.05$.

These data indicate that rate variation and stress variation both influence a listener's sensitivity to the spectral information for vowels. It appears, however, that their influences are not commensurate; an increase in rate and a decrease in stress invite qualitatively different vowel identification patterns. This conclusion is compatible with a growing body of acoustic data (Verbrugge & Shankweiler, 1977; Harris, 1978; Gay, 1978), and with production data. Harris (1978) has found, for example, that electromyographic activity associated with tongue body movement varies principally in terms of duration for rate changes and in terms of amplitude for stress changes.

The rate effects observed in this study deserve special mention. We have argued elsewhere (Verbrugge, Strange, Shankweiler, & Edman, 1976; Verbrugge & Shankweiler, 1977) that rate information serves principally to aid a listener in adjusting to variations in vowel duration. This observation has been based on the fact that when syllables were excised from sentence carriers and presented in citation form, or were swapped into sentence carriers spoken at different rates, errors tended to increase for short-long vowel pairs. However, in the current study, the interaction of rate and vowel duration was controlled through the synthesis of syllables with vowel durations that matched those of the natural syllables produced in each carrier. Nonetheless, rate effects were observed, suggesting that rate aids a listener in adjusting not only to variations in vowel duration, but to spectral variations as well.

Finally, these data can be at least partially addressed in terms of center formant frequency measures made for the natural syllables produced in the carriers. Formants in the natural slow-destressed syllables were reduced in vowel color relative to the slow-stressed syllables, particularly formant one (see Figure 1). Given this circumstance, it is not surprising that when the stressed formants were synthetically paired with destressed prosody, an increase of /ɛ-æ/ and /ʌ-ɑ/ confusions resulted.

A look at the formant space for the fast-stressed syllables is somewhat less helpful. The formants of fast-stressed syllables (Figure 2) do differ significantly from those of slow-destressed syllables (Figure 1), as would be expected given the results of this study. However, fast-stressed formants do not differ appreciably from slow-stressed formants (Figure 2), making an account of the rate effect observed here unclear. Some data reported by Gay (1974) may speak to the rate effects, however. Averaging over a number of tokens of the point vowels /i, a, u/ spoken by several talkers, Gay found that the vowel space for fast speech differed from that for slow speech in that first formant frequencies were higher for all vowels. Given this circumstance, one would predict a predominance of long vowel-short vowel confusions when slow-syllable formants are paired with fast prosody, and this is what we observed.

In conclusion, we have shown that the spectral information for vowels is perceived differently as a function of rate variation and stress variation. These distinct alterations in perceptual sensitivity correspond, at least in part, to the differing effects that these two transformations have on the spectral properties of vowels. Thus, the perceptual results provide indirect support for the hypothesis that rate and stress are independently regulated in production.

REFERENCES

- Gay, T. A cinefluorographic study of vowel production. Journal of Phonetics, 1974, 2, 255-266.
- Gay, T. Effect of speaking rate on vowel formant movements. Journal of the Acoustical Society of America, 1978, 63, 223-230.
- Harris, K. S. Vowel duration change and its underlying physiological mechanisms. Language and Speech, 1978, 21, 354-361.

- Lindblom, B. E. F. Spectrographic study of vowel reduction. Journal of the Acoustical Society of America, 1963, 35, 1773-1781.
- Peterson, G. E., & Barney, H. I. Control methods used in a study of the vowels. Journal of the Acoustical Society of America, 1952, 24, 175-184.
- Verbrugge, R. R., & Shankweiler, D. P. Prosodic information for vowel identity. Haskins Laboratories Status Report on Speech Research, 1977, SR-51/52, 27-35.
- Verbrugge, R. R., Strange, W., Shankweiler, D. P., & Edman, T. R. What information enables a listener to map a talker's vowel space? Journal of the Acoustical Society of America, 1976, 60, 198-212.

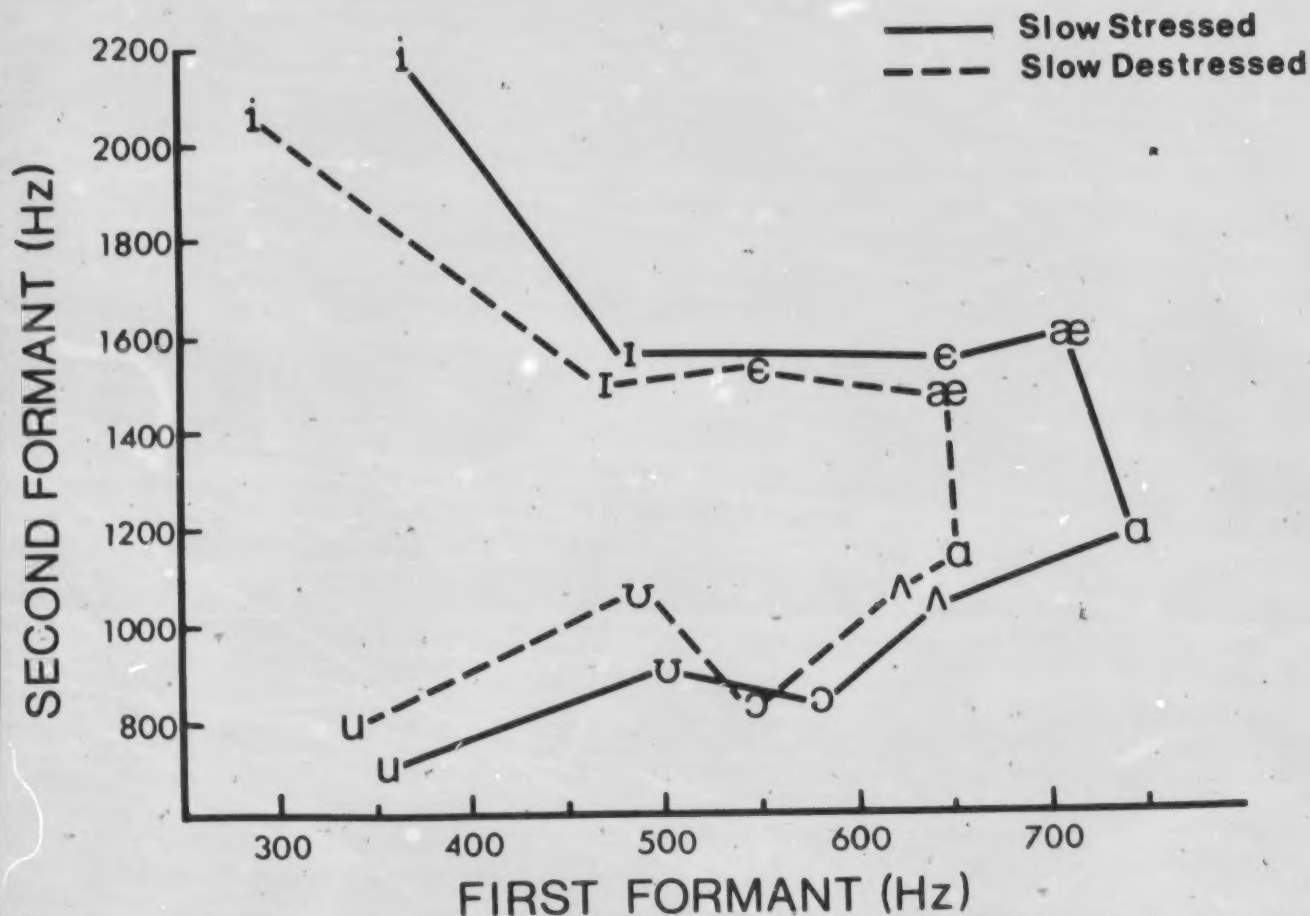


Figure 1. Center frequencies of first and second formants of nine vowels spoken in /pVp/ syllables: slow-stressed and slow-destressed conditions.

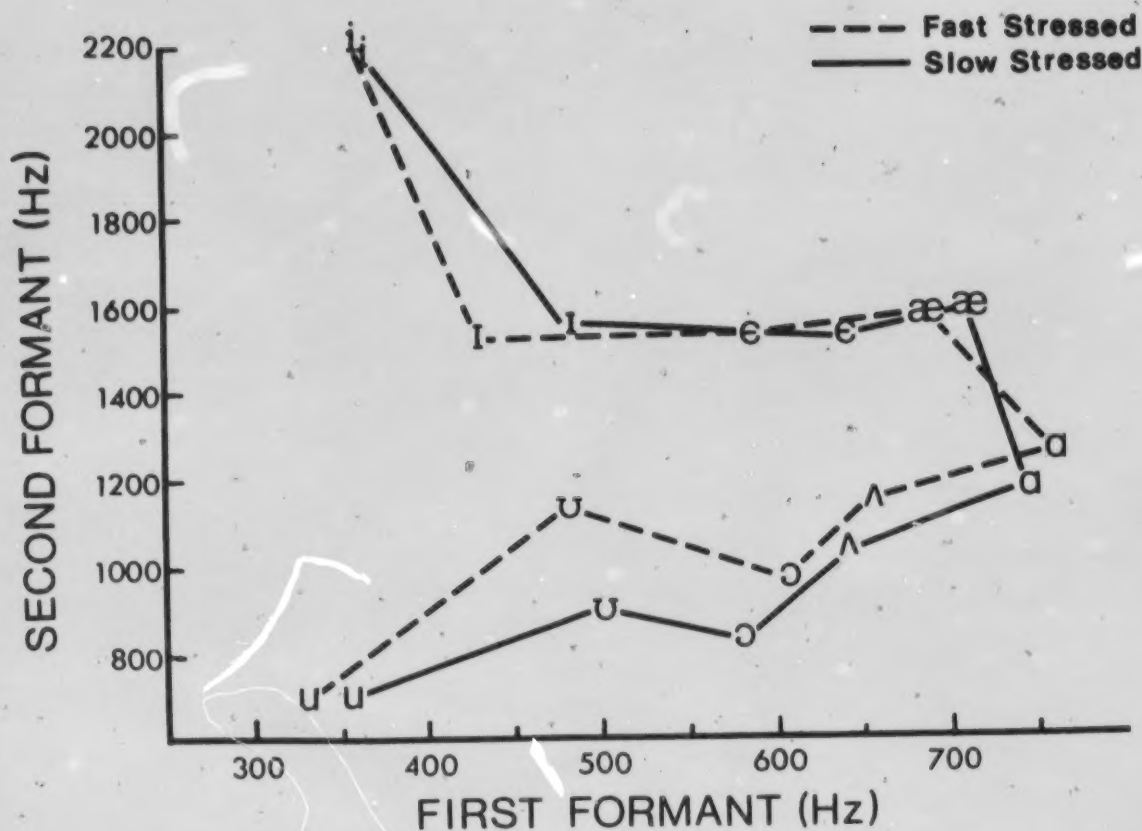


Figure 2. Center frequencies of first and second formants of nine vowels spoken in /pVp/ syllables: slow-stressed and fast-stressed conditions.

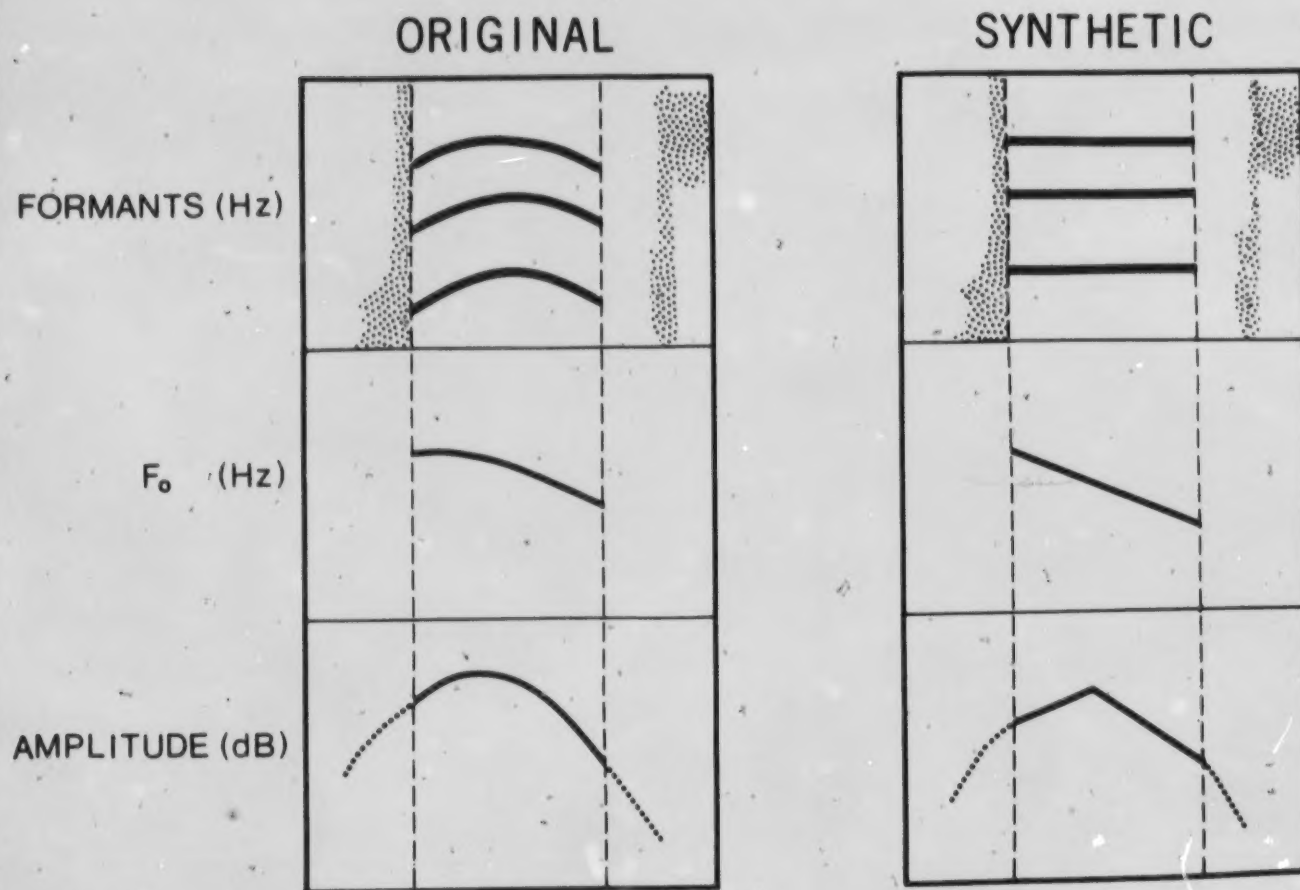


Figure 3. Comparison of a natural /pVp/ syllable with a corresponding syllable in which the vocalic region has been synthesized.

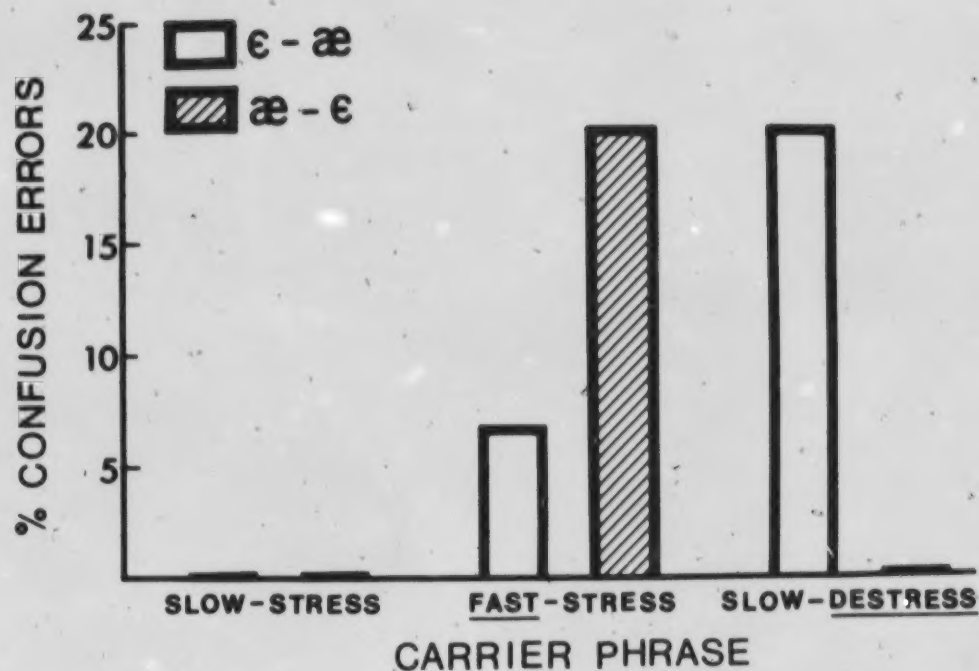


Figure 4. Listener confusions of the vowels /ε/ and /æ/ for /pVp/ syllables with synthetic vocalic regions in which the formants of slow-stressed syllables were paired with the prosody of either slow-stressed, fast-stressed, or slow-destressed carriers.

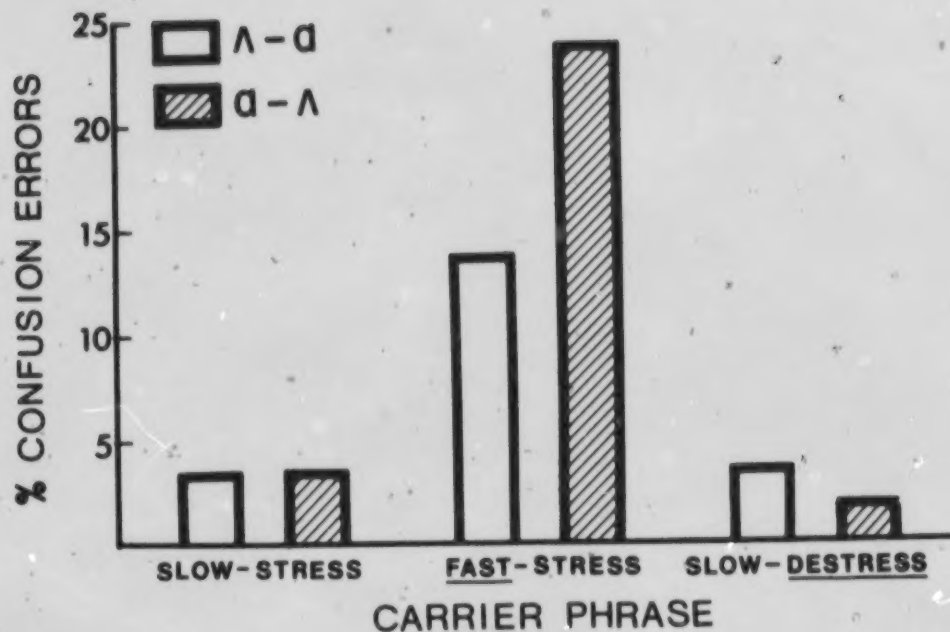


Figure 5. Listener confusions of the vowels /Λ/ and /a/ for /pVp/ syllables with synthetic vocalic regions in which the formants of slow-stressed syllables were paired with the prosody of either slow-stressed, fast-stressed, or slow-destressed carriers.

SONORITY AND SYLLABICITY: ACOUSTIC CORRELATES OF PERCEPTION

Patti Jo Price*

Abstract. Evidence is accumulating that native speaker-hearers are not as consistent, confident, or in agreement on counting the number of syllables in natural utterances as is commonly assumed (Lebrun, 1966; Bell, 1975; Price, Note 1). There are, however, instances where speaker-hearers give clear, consistent syllable counts. It is the position of this paper that the unclear cases as well as the clear cases are phonetically classifiable in terms of sonority. The experiments presented here are intended to delimit what is meant by sonority in acoustic terms.

The syllable is more often appealed to than defined (see surveys in Bell, 1978; Price, Note 1). The problems arising from attempted definitions are sometimes "explained away" by positing the syllable as a "natural perceptual unit." In this view, native speakers have strong intuitions about syllables, but definitions cannot be developed from these intuitions due to complex interactions of morphology, phonology, orthography and phonetics. However, evidence is accumulating that even this weak claim for the syllable may not hold. Bell (1975) tried a variety of methods in attempting to elicit natural units. Lebrun (1966) asked for syllable counts of short sentences repeated as often as subjects wished. Price (Note 1) asked for syllable counts of very short utterances with dialect background strictly controlled. In all these studies, where the assumption about the intuitive status of the syllable was tested, the results converged: native speaker-hearers were not extremely consistent, confident or in agreement on syllable counts. Moreover, automatic segmenting algorithms tend to fail in areas phonetically similar to those where native speakers are inconsistent or disagree: neighboring segments of roughly the same degree of sonority. Mermelstein (1975) mentions cases of syllabic vs. non-syllabic /r/ or /n/ ("horizon" as /hrajzen/ or "apparently" as /æppɛrnlɪ/ [sic]) and contiguous vowels as in "so I." The inconsistency of listeners, the lack of agreement across listeners, and the failure of algorithms to match dictionary syllabifications do not necessarily imply that the syllable does not exist or is useless. Clear cases exist, and, further, the unclear cases may be taken as evidence that a more flexible definition of the syllable is necessary, i.e., a definition that accounts for both the clear and the unclear cases. Such a definition in terms of sonority peaks will be outlined here. The experiments reported here investigate the acoustic correlates of the perceptual term "sonority."

*Also University of Pennsylvania.

Acknowledgment. I wish to thank Len Katz, Andrea Levitt, Leigh Lisker, and Michael Studdert-Kennedy for discussion, criticism and many helpful comments; and BRSC Grant RR-05596 to Haskins Laboratories for support.

The terms "prominence" or "sonority" have been applied to various aspects of speech: as an overall feature of voice quality (see, e.g., Wedin, Leanderson, & Wedin, 1978), as a feature of stress or accent carried by certain syllables (see, among many others, Gaitenby & Mermelstein, 1977), and as a feature of segments forming the internal structure of syllables. Only the latter usage of the terms will be dealt with in the present study. Acoustic correlates for these terms that have been investigated, however, show a great deal of overlap. Most of the studies consider fundamental frequency (absolute value or extent of excursion), intensity, and duration. The relative roles of these acoustic attributes are under debate, perhaps largely because of methodological differences. I know of no studies that have investigated experimentally the acoustic features of sonority for segments smaller than the syllable, although there has been a fair amount of theorizing.

Since Sievers (1893) the internal structure of the syllable has been discussed in terms of sonority, strength or prominence (see, for example, Bloomfield, 1933, p. 120; Venneman, 1972; Hooper, 1976). Bell and Hooper (1978) provide a good survey of cross-linguistic evidence for such hierarchies. The basic notion is that syllabic peaks are peaks of sonority, and that segments increase in sonority before the peak and decrease in sonority after it. This implies that English /l/ is more sonorant than English /p/, because instances of /#plV/ and /_Vlp#/ occur but not /#lpV/ or /_Vpl#/. The two latter examples may in fact be realized in English, but only if the /l/ is sonorant enough to be a syllable peak itself, as in "I'll put it away" (/lpUDIDaweɪ/), or "people" (/pipl/). The cross-linguistic evidence may be taken to mean that sonority hierarchies are language specific: Russians seem to feel that /#rtV/ structures are one syllable, although in English a monosyllable of this structure is impossible. If utterance-initial Russian /rtV/ structures are monosyllabic structures, then, in terms of sonority, we are forced to say that Russian /t/ is more sonorant than Russian /r/, and we cannot say the same for English. However, this does not necessarily mean that Russian and English /r/s are of the same sonority, or that sonority has no explanatory value. If an acoustic definition of sonority is developed, the relative sonority of linguistic units from different languages can be compared without reference to the language specific phonotactics of the two linguistic systems.

A fundamental generalization of the sonority theory is that vowels are more sonorant than consonants. A crucial aspect of the theory, however, is that it is useful to divide the set of phonetic segments into a richer classification system than one involving only consonants and vowels. Even linguists who make no specific mention of sonority hierarchies may define the syllable in terms of a vocalic nucleus surrounded by consonantal margins (onset and/or coda) (Hjelmslev, 1938; Trager & Bloch, 1941; Trager & Smith, 1951; Hockett, 1958; Greenberg, 1962; Chomsky & Halle, 1968; Studdert-Kennedy, 1976). Peaks of sonority are, in general, vowels; the troughs are generally consonants. Defining syllables in terms of alternations of consonants and vowels works insofar as the classes of consonants and vowels are clear. By examining the cases of clear and unclear vowels, we can outline the classes of clear and unclear sonority peaks, and, hence, predict listener inconsistency, disagreement, or possible problems for automatic segmentation algorithms and the intuitions of a native speaker-hearer.

"Clear," "good" or "prototypical" vowels correspond to prototypical syllabic peaks and are fairly easy to describe in articulatory or in acoustic terms. They are characterized by an open vocal tract, vibrating vocal folds, and relatively long duration. Good consonants or syllabic margins are characterized by the opposite: a constricted or closed vocal tract, interrupted voicing, and relatively short duration (a "transient" as opposed to "steady-state" character of the more audible portions of the articulation). These three factors--degree of opening of the vocal tract (OPENING), glottal source characteristics (VOICING), degree of transience (RATE OF CHANGE)--are all involved in sonority. All the experiments to be reported in this paper bear on the meaning of sonority and the role it plays in syllabicity judgments. In the acoustic domain, these three factors may correspond to the presence vs. absence of a clear formant structure, voice vs. hiss (or no) excitation source, and steady-state vs. transient formant patterns. The RATE OF CHANGE characteristic may apply to parameters other than formant structure (fundamental frequency or amplitude, for example), but other aspects will not be specifically investigated here. In the idealized situation, then, chains of syllables are series of vocal tract openings and closings with the open parts (syllabic peaks) corresponding to vowels and the closed parts corresponding to consonants. In the clearest cases this is true, with the exceptions that: (1) there is a tendency to think of the closing and/or opening gesture as constituting the consonantal component (rather than the "closed" part itself), and (2) there is a tendency to think of these opening and closing gestures as organized into discrete consonant-plus-vowel units.

It is important to notice that the characteristics OPENING, VOICING, and RATE OF CHANGE are all relative rather than absolute terms. Furthermore, there are many cases where only one or two of these qualities may appear. For example, insofar as openness of the vocal tract indicates degree of vocalicness, open vowels (say, [a]) are more vowel-like than close vowels (say, [i] or [u]). A number of linguists (e.g., Hockett, 1942; Pike, 1943, pp. 110-111; Jones, 1950, p. 15) treat [j] and [w] as non-syllabic counterparts of [i] and [u]. The orthography chosen may highlight this view: they are often written identically, sometimes with a diacritic added to distinguish them. Close vowels are not only similar to glides (sometimes called "semi-vowels" or "semi-consonants"), but they also risk confusion with segments that are not "semi" but "real" consonants: slight deviations in control of air supply for constricted vowels can produce friction noise, causing a similarity to fricatives (as in, for example, American English "heed your" [hidʒr]--[hidʒr]). Some vowels are more vowel-like than others with respect to OPENING, VOICING, or RATE OF CHANGE. All three characteristics are a matter of degree. Voicing is a matter of degree both in its relative onset (see, e.g., Lisker & Abramson, 1964) and in the amount of accompanying friction noise (as in, for example, voiced vs. murmured vs. whispered vowels).

Furthermore, these three characteristics are relatively independent. That is, they may differ as indices of how vocalic (sonorant) a particular segment is. For example, in voiceless vowels, the mouth can be very open, steady state portions may be clearly present, but voicing is absent. Glides represent a case in which the vocal tract is relatively open and voicing is present, but there is a rapid rate of change. There are also cases in which a voiced steady state period occurs when the vocal tract is obstructed, as for nasals, voiced obstruents, and liquids. Voiceless fricatives may have a long

steady state period, but rank low on the scales of VOICING and OPENING. In fact, the set of clear vowels or clear consonants is probably smaller than the set of unclear cases.

When one considers the combinatorial properties of these elements, the problem of syllables becomes more complex. In terms of the characteristics of prototypical vowels (OPENING, VOICING, and RATE OF CHANGE), prototypic syllables can be defined as alternations of prototypic vowels and prototypic consonants. This predicts that listeners will agree more on the number of syllables in utterances that consist of alternations of prototypic consonants and vowels than they will on alternations of the less clear cases. Support for this hypothesis is found in Price (Note 1).

The present study considers liquids (English /r/s and /l/s) in /C V/ position. The degree of openness of the vocal tract cannot be systematically varied for most sounds, since we tend to define classes of phones largely with respect to this aspect. It is possible to vary relative and absolute duration, amplitude, and voice onset time. The present study investigates these aspects of sonority in the case of the pairs plight-polite and prayed-parade. There are many such pairs where the sonority or prominence of the /l/ or /r/ may be all that is needed to keep lexical items distinct: bray-beret, round-around, long-along, set lit-settle it. There are also more ambiguous pairs where it is not clear that there is a distinction at all: hire-higher, aisle-I'll, etc. Assuming that syllabic peaks are peaks of sonority, then increasing the sonority of certain segments of variable sonority should lead to an increase in the number of syllables perceived. Experiment 1 tests the roles of duration and amplitude for their contribution to the sonority of /r/ in natural productions of prayed and parade. Experiment 2 tests voice onset time and the relative roles of voicing, hiss, and silence in a synthetic plight-polite continuum. Experiment 3 tests the roles of relative vs. absolute /l/ durations in the same synthetic plight-polite continuum.

Relative intensity and relative duration together led to the best prediction of perceived syllable stress in Gaitenby and Mermelstein's (1977) study, with the value of intensity outweighing that of duration and of fundamental frequency. One might expect some overlap in perception of prominence within and across syllables, but the two are not necessarily identical.

EXPERIMENT 1

/r/ Duration vs. /r/ Amplitude in Natural Productions of prayed and parade

In this experiment, amplitude and duration of the /r/ portions (defined as the portions where F2 and F3 are close to each other, an acoustic indication of retroflexion) in natural productions of prayed and parade were manipulated by computer editing and presented to naive listeners for labeling.

Measurement Data

Ten productions each of prayed and parade by each of two talkers, one male and one female, were measured. Voice onset time was measured from waveforms, /r/ duration from spectral displays. Amplitude of aspiration and of the /r/ were measured in dB down from the peak by computer analysis.

While the duration of aspiration (VOT) was, on the average, about 10 msec longer for prayeds than for parades, the range of these durations for parade (40-60 msec) was wholly included in the range for prayed (40-80 msec). Thus, it was decided not to manipulate this parameter in the present experiment. The amplitudes of aspiration did not differ significantly either in range or in mean value. /r/ durations did vary significantly: the mean for prayed was found to be 80 msec with a 55 to 110 msec range, and the mean for parade, 145 msec with a range of 110 to 170 msec. While some tokens of parade were apparently pronounced by the male talker as /pəreyd/, as evidenced by spectral displays, amplitude envelopes (two humps in the display), and by listening, all productions by the female talker (and most productions by the male talker) were pronounced /preyd/--with syllabic /r/. The amplitude levels for the /r/s were measured at the /r/ peak in dB down from the peak for that token, where such a peak occurred. Where no such peak occurred, these levels were averaged over 12.6 msec intervals throughout the /r/. This ad hoc procedure may not result in a meaningful measurement. In fact, average amplitude levels by this measure differed by only 1 dB.

Stimuli

In this experiment /r/ duration and amplitude were altered independently. Based on the measurement data, an "average" token for each of the source words was selected from among the productions by the male talker, since his formants were easier to track and measure, and the longer pitch periods made it easier to extend the /r/ by pitch pulse iteration without disturbing the naturalness of the tokens. From the parade chosen, /r/ duration was shortened by deleting pitch pulses after onset of voicing. Deleting 30 msec resulted in a derived stimulus with a 115 msec /r/. Similarly, deleting 60 msec resulted in a derived stimulus with an 85 msec /r/. The most drastically shortened stimulus, then, had an /r/ duration roughly equal to the mean for the set of prayeds. Amplitude of the /r/ portion was decreased by 6 dB for the original and for the portion of the /r/ remaining in the shortened versions. For the prayed chosen, /r/ duration was increased by iteration of the first pitch pulse. Two stimuli were derived in this fashion, one with /r/ duration increased by 30 msec (110 msec total /r/), and one with /r/ duration increased by 60 msec (140 msec total /r/). Adding 60 msec to the original token created a stimulus whose /r/ duration (140 msec) was roughly equal to the mean value for tokens of parade (145 msec). For each of these three /r/ durations two amplitude levels were used: the original, and 6 db up from the original. Four tokens of each stimulus appeared in a randomized test sequence on separate tapes for each source word.

Subjects

The subjects were 12 paid volunteers (all Yale undergraduates) who were asked to listen to the tapes twice, in counterbalanced order, once to count

/r/ DURATION & AMPLITUDE N=96

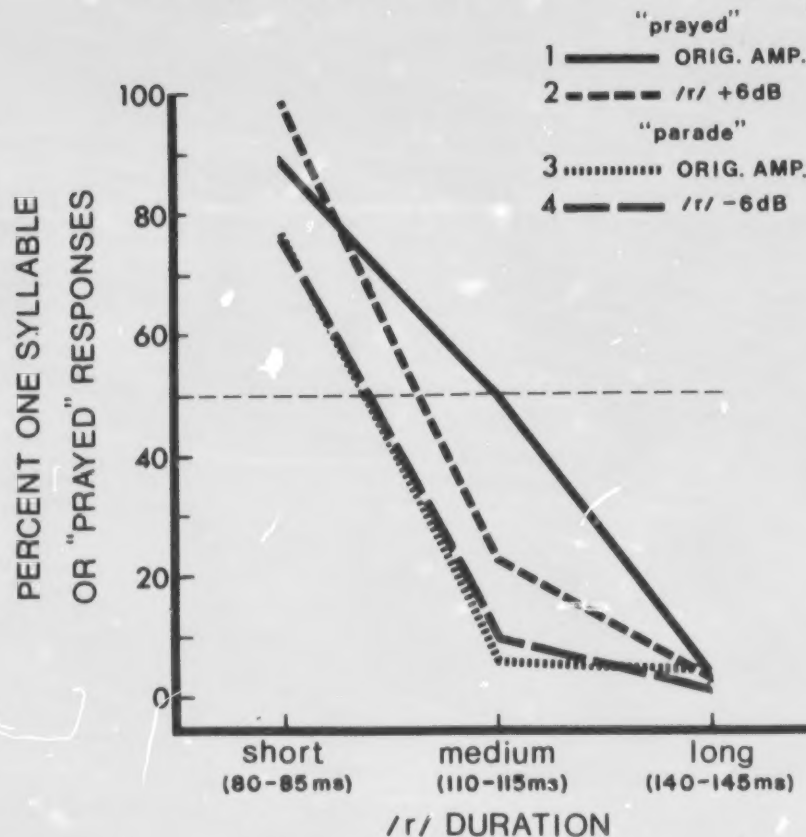


Figure 1. N = 96. "Short," "medium," and "long" refer to /r/ durations. The "short" condition corresponds to prayeds with original /r/ duration and to parades with 60 msec of the /r/ deleted. The "medium" condition corresponds to /r/ lengthened or shortened by 30 msec for sources prayed and parade respectively. The "long" condition refers to parade of original /r/ duration and to prayed with /r/ duration increased by 60 msec. Represented in plots 1 and 2 are "prayed" responses to stimuli derived from source prayed. Plots 3 and 4 correspond to source parade. Plots 1 and 3 correspond to original /r/ amplitude levels, plots 2 and 4 to manipulated /r/ amplitude levels. Note that /r/ duration has a decisive effect on labelings. Amplitude may have some effect for the "medium" condition.

syllables and once to identify words by circling either "prayed" or "parade" on prepared answer sheets.

Results and Discussion

The responses resulting from these two tasks (syllable counting and word identification) did not differ significantly: "one syllable" responses correspond to "prayed" responses to within 6 percent for every source stimulus. The two tasks combined yield an $N = 96$. Figure 1 plots percent one syllable or "prayed" responses vs. /r/ duration for both source words, and for both amplitude levels. It is clear from this figure that /r/ duration has a decisive effect on listener judgments for both source words. It appears that the effects of amplitude are negligible for short and for long /r/ durations. The case of the "medium" durations is less simple: for source prayed, amplitude affected judgments significantly, but this was not the case for source parade. This asymmetry may indicate a general difference between productions of prayed and productions of parade, at least for this talker, or it may be due to token-specific differences. Although an "average" token of each source word was chosen, these were natural productions and, hence, differ along many uncontrolled dimensions. In any case, the set of "medium" duration stimuli support the conclusion that duration is a more effective cue than amplitude: when the duration used resulted in judgments split between the two words (curve 1), amplitude had a significant effect (curve 2); when, however, the duration used resulted in judgments strongly in favor of one word or the other ("parade," in this case, curve 3), amplitude had little effect (curve 4). The measurement data indicate that absolute /r/ durations may well be ambiguous as indices of prayed vs. parade in natural productions: the longest /r/ of prayed was of the same duration as the shortest /r/ for parade. When the words are embedded in sentences, it is likely that the ranges of the /r/ durations for the two source words will overlap. In sum, duration of /r/ for these words seems to be a sufficient cue to their distinction. Amplitude may play a role where this cue is neutralized. While more open vowels are generally louder (of higher amplitude level) than less open vowels, differences in formant frequencies, or vowel color, are also generally involved. If sonority is considered in articulatory terms, then the rather small effect of amplitude is reasonable, given that spectral information was unchanged. The independent testing of amplitude and spectral information as they relate to the openness of the vocal tract is left for future research.

EXPERIMENT 2

Voice Onset Time in Synthetic Stimuli

The measurement data for the prayeds and parades revealed consistently longer mean values of voice onset time (VOT) for prayeds than for parades for both talkers. Although the ranges of these values overlap heavily for both talkers, the mean difference is 10 msec for the male talker, 20 msec for the female. Further, the longer VOTs are correlated with shorter /r/ durations. The situation is not entirely parallel to that of voiced vs. voiceless stops in initial position: (1) the duration of the segment following the initial stop (as well as VOT value) serves to distinguish "prayed"-"parade," "plight"-"polite," etc., but not initial /bdg/ vs. /ptk/; (2) the differences in VOT

correspond to differences not in the voicing of the initial stop but in the syllabicity of the following segment, and (3) increases in VOT are not necessarily correlated with increases in formant frequency onset values, since the liquid may be steady state throughout a wide range of VOT values. However, both situations involve coordination of vocal tract opening and the onset of voicing. In other words, sonority is not merely a matter of opening and closing the vocal tract, but of vocal tract dynamics and their interaction with laryngeal control. Thus, a continuum that switches judgments from one to two syllables based on VOT alone is evidence that the perceptual significance of the relative timing of vocal tract gesture and laryngeal pulsing, as evidenced by VOT, is generalizable beyond the class of initial stop consonants.

Stimuli and Subjects

The stimuli for this experiment were prepared on the OVE-3 synthesizer at Haskins Laboratories. Stimuli perceived as "plight" and "polite" were created. Spectrograms of endpoint stimuli are shown in Figure 2. In order to avoid the intrusion of initial /b/ percepts, the shortest VOT used was 49 msec. VOT was increased from 49 msec to 126 msec in 7 msec steps. The increase in VOT decreased the buzz-excited steady state /l/ duration from 91 msec to 14 msec, thereby increasing the hiss-excited steady state /l/ from 14 to 91 msec. A similar set of stimuli was created in which silence replaced the hiss between initial burst and voicing onset. These stimuli are probably less representative of actual articulations than the first set, but they do permit the investigation of the perceptual effect of hiss vs. silence. It is reasonable to use these stimuli a priori since aspiration hiss may not always be audible in speech contexts, and a posteriori because they in fact result in a convincing "plight"-"polite" continuum. Four randomizations of these 24 stimuli were presented to 13 paid volunteers (Yale undergraduates) for labeling as "polite" or "plight." The graphs in Figure 3 thus represent N = 48.

Results and Discussion

In Figure 3 it is seen that VOT is an effective cue to the plight-polite distinction. Further, it appears to make a difference whether the period between burst and onset of voicing is noise-filled or silent: subjects, on the average, need about 16 msec longer voiced steady state /l/ to hear polite vs. plight when silence replaces hiss in this interval. The steady state portion of the /l/ is crucial to hearing polite vs. plight, but the voiced part is more critical than the voiceless part. That is, the total steady state /l/ is not the critical factor here: all stimuli have the same duration in this respect. What appears to be critical is the overall sonority of the /l/. As is shown here, duration of voicing of the /l/ effectively switches judgments from "plight" to "polite" in both the hiss and the silent conditions. If, however, hiss is present between burst and voicing onset, the cross-over is realized with a shorter voiced /l/ duration (about 16 msec shorter) than if this interval is silent. This suggests a sonority hierarchy of voicing over hiss over silence.

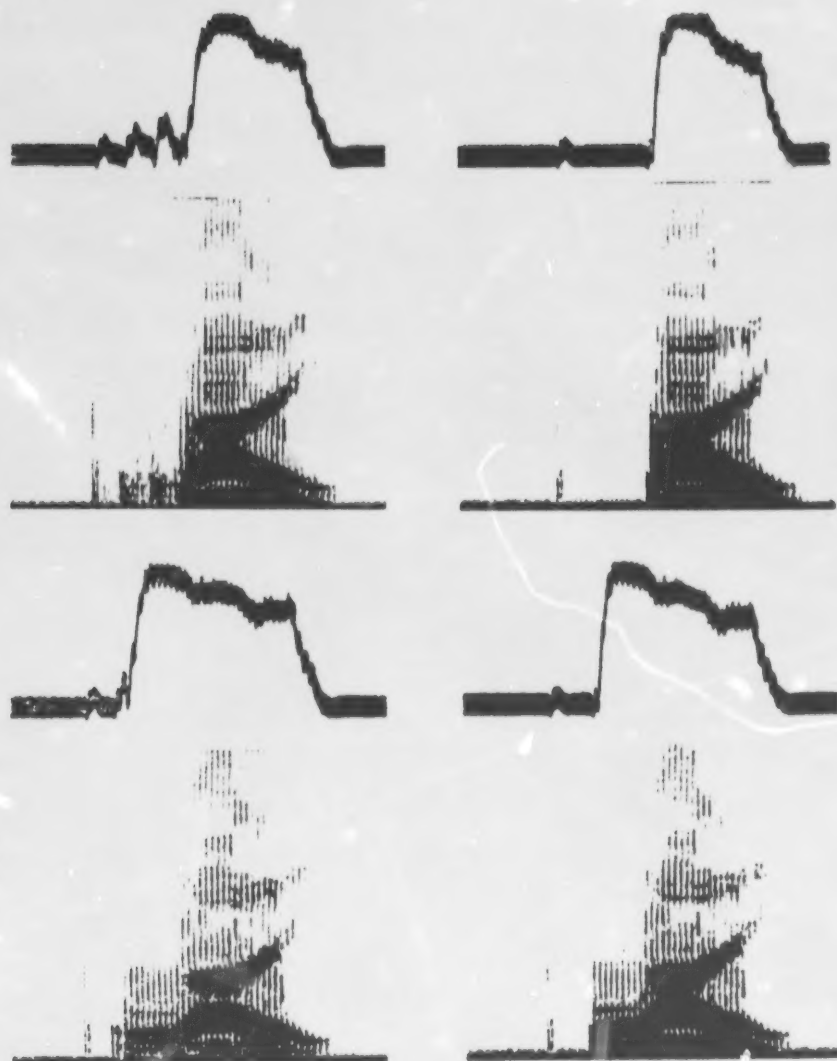


Figure 2. SPECTROGRAMS OF SYNTHETIC STIMULI. At the left are the endpoints of the hiss condition of the synthetic "plight"-"polite" continuum. At the right are the endpoints for the silence condition of this continuum. The two displays at the top represent the stimuli with longest VOT values, and, hence, shortest voiced /l/ duration. The displays at the bottom represent the shortest VOT values used, which correspond to the longest duration of the voiced steady state /l/.

VOT OR DURATION OF VOICED /l/, N = 48

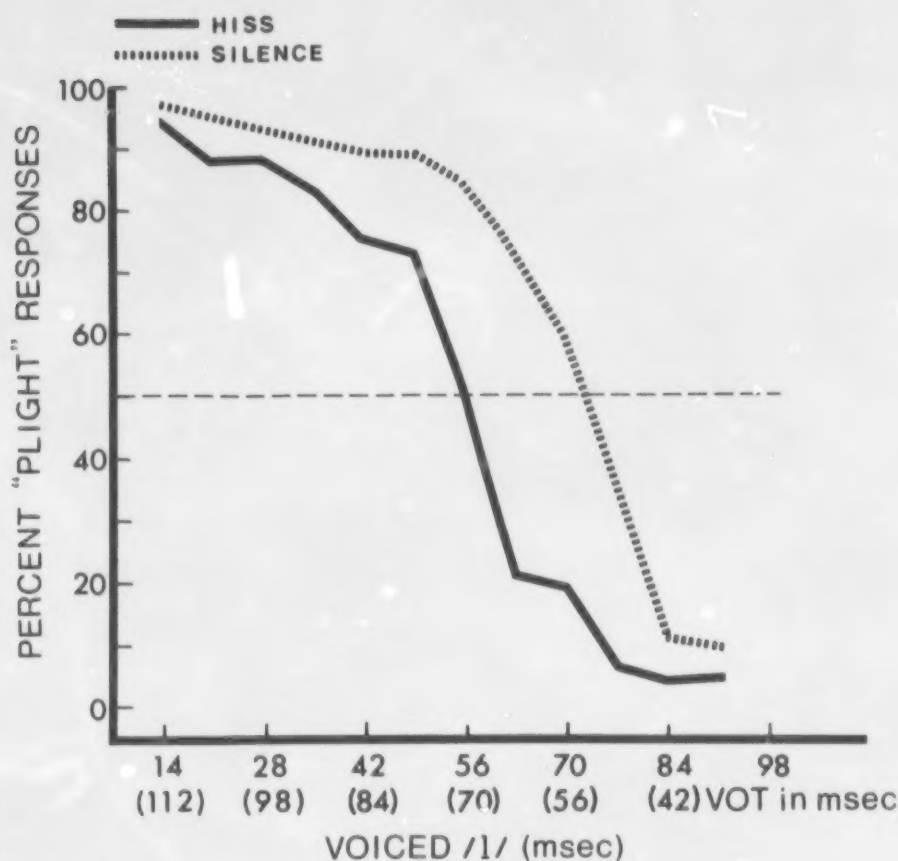


Figure 3. N = 48. The solid line indicates percent "plight" responses to stimuli in which the interval between burst and voicing onset was hiss filled. The dotted line represents the condition in which silence filled this interval. The abscissa is labeled with duration of voiced /l/. Underneath the /l/ duration figures, the corresponding VOT values appear in parentheses. All stimuli have the same steady state /l/ duration (91 msec); they differ in the point at which formant excitation is switched from hiss to buzz. Note that the longer the voicing of the /l/ (i.e., the shorter the VOT), the more "polite" responses elicited (i.e., fewer "plight" responses). Further, the duration of voicing of the /l/ at the "plight"-"polite" cross-over point is about 16 msec longer when silence rather than hiss is present in the interval between burst and voicing onset. This suggests that, with respect to duration, voicing is more effective than hiss in cueing "polite" rather than "plight," and that hiss, in turn, is more effective than silence.

EXPERIMENT 3

Relative vs. Absolute Duration in Synthetic Stimuli

Stimuli and Subjects

The third and final experiment to be reported here involves the issue of rate, or absolute vs. relative durations. In this experiment two factors are pitted against each other: the absolute duration of the steady state /l/ and the overall rates of the stimuli. Stimuli similar to those used in the hiss condition of Experiment 2, but with larger step sizes, were used in this experiment under four conditions:

- (1) ORIGINAL: stimuli of Experiment 2 (hiss condition) with VOT varied from 42 to 126 in 14 msec steps (voiced /l/ duration thereby decreasing from 84 to 0 msec);
- (2) EXTENDED: /l/ duration of the stimuli increased by 35 msec, thus adding two new stimuli;
- (3) FAST: stimuli of condition (1) played out at a 40% faster rate, and
- (4) EXTENDED FAST: stimuli of condition (2) played out at a 40% faster rate.

Conditions (1) and (4) thus represent stimuli with the same absolute /l/ duration (90 msec), but played out at different rates. Conditions (1) and (3), on the other hand, have the same /l/ durations relative to the duration of the entire stimulus. Likewise, stimuli in conditions (2) and (4) have /l/ durations of the same percentage of overall duration, though the two sets of stimuli are played out at different rates. Figure 4 shows spectrograms of the stimuli used for the shortest VOT value in the EXTENDED and EXTENDED FAST conditions. The step sizes were 10 msec in the fast conditions, 14 msec in the other two conditions. The extended conditions thus involved two more stimuli than the unextended conditions. Three tokens of the stimuli were presented to 10 paid volunteers (Yale undergraduates): $N = 30$.

Results and Discussion

Figures 5a and 5b show "plight" responses as a function of duration of voiced /l/ expressed as a percentage of overall duration. VOT and voiced /l/ duration are here inversely correlated: longer VOT values correspond to shorter voiced /l/ durations. Note that for both the original (Figure 5a) and the extended (Figure 5b) stimuli, an increase in overall rate elicits more "plight" responses. That is, when relative durations are equated, an increase in rate does affect listener judgments. The effect of absolute duration is shown in Figure 6. In this figure "plight" responses are plotted as a function of the absolute duration of voiced /l/. It thus appears that, other things being equal, the absolute duration of the voiced portion of the /l/ has a greater effect on listener judgments than its relative duration, at least for the rates and the durations used here. Further research may reveal that for certain ranges of absolute duration, the relative duration of a segment



Figure 4. SPECTROGRAMS OF SYNTHETIC STIMULI USED IN EXPERIMENT 3. These displays represent the shortest VOT (longest voiced /l/) stimuli used for the EXTENDED, and the EXTENDED FAST conditions.

RELATIVE DURATION OF VOICED /l/

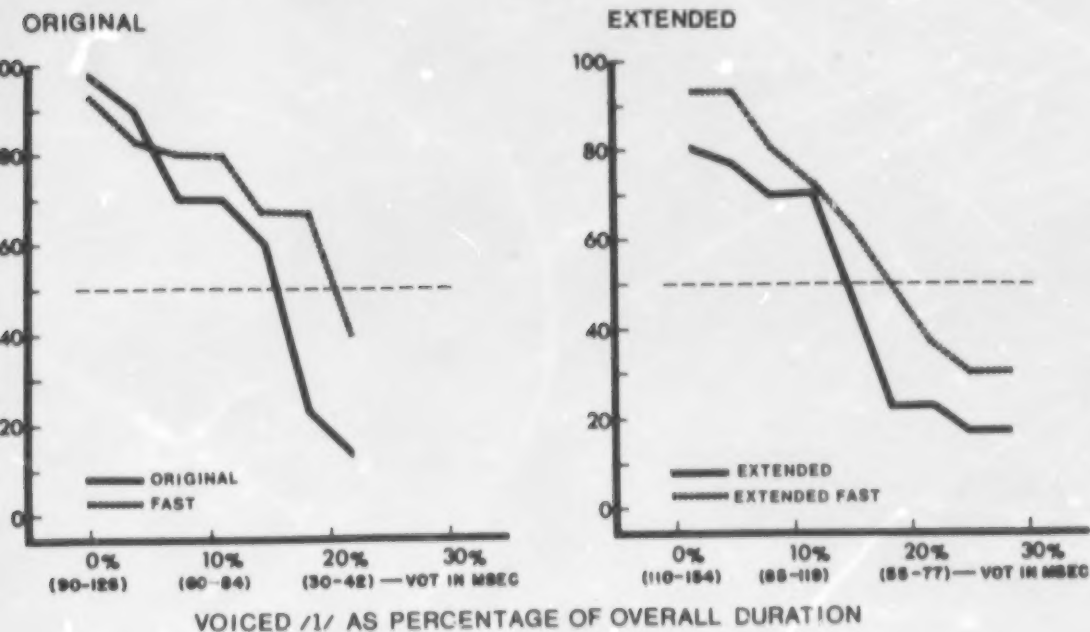


Figure 5. Figure 5a (left) and Figure 5b (right). $N = 44$. Plotted here are "plight" responses as a function of voiced /l/ duration. (VOT values are in parentheses.) /l/ durations are expressed here in percentages of overall stimulus duration. Note that for the same relative durations (either for original or extended stimuli), the faster overall rates elicit more "plight" responses.

ABSOLUTE DURATION OF VOICED /l/

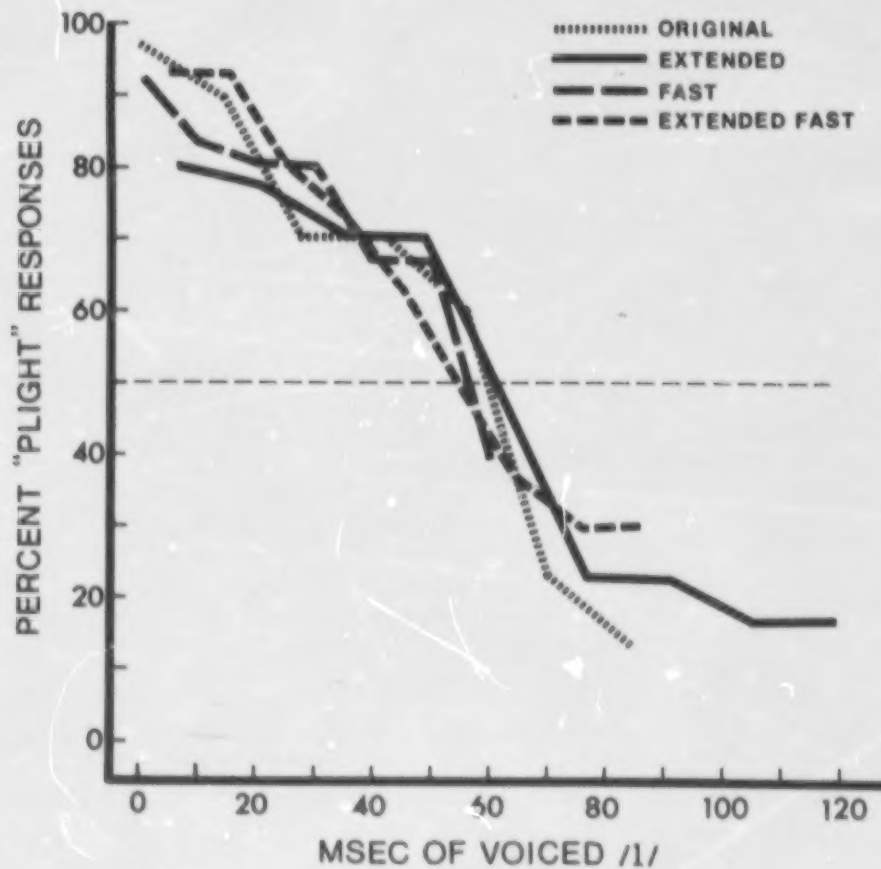


Figure 6. $N = 44$. The same data presented in Figure 5 are plotted here as a function of absolute /l/ duration. As is seen in this figure, for the /l/ durations and rates used, absolute rather than relative duration seems to be the crucial factor in listener judgments. The cross-over point here is 55 to 65 msec voiced /l/. This is consistent with the value found in Experiment 2.

with respect to its surround makes a difference, but these data so far indicate a different story: that absolute duration is a more effective indicator of sonority than is relative duration.

CONCLUSIONS

If it is assumed that the auditory term "sonority" can account for the syllabic vs. non-syllabic distinction of certain segments, then this study provides evidence for acoustic correlates of this term. Sonority, like other perceptual terms such as "pitch" or "loudness," has multidimensional acoustic correlates. The experiments presented here bear on the roles of duration, amplitude, voicing, hiss, and silence as they relate to sonority. The results of these experiments support the following hypotheses:

- (1) duration is a more effective cue to sonority than is amplitude,
- (2) amplitude may play a role when duration is ambiguous,
- (3) when duration is manipulated, voiced segments tend to be more sonorant than hiss-excited segments, which in turn appear more sonorant than silence,
- (4) absolute duration is more important to perceived sonority than relative duration.

Acoustic or auditory correlates have been proposed (but not tested) for the perception of syllabic peaks vs. margins. Fischer-Jørgensen (1975) suggested that liquids are auditorily weaker than vowels since most of their energy is concentrated in the first formant. Fant (1969/1973) suggested a weighted sum of the intensities of F1 and F2 compared to that in adjoining segments. Gaitenby and Mermelstein's (1977) weighting function, which favors the frequencies between 500 and 4,000 Hz, implies a similar acoustic-auditory emphasis, although in this case the weighting is done in order to analyze syllabic stress rather than internal syllable structure. Left for further research is the implementation of these suggestions in a test of the hypotheses proposed in the present study.

REFERENCE NOTE

1. Price, P. The Syllable. Unpublished M.A. paper, University of Pennsylvania, 1978.

REFERENCES

- Bell, A. If native speakers can't count syllables, what can they do? Bloomington, Ind.: Indiana University Linguistics Club, 1975.
- Bell, A. Segment organization phenomena and their explanations. In A. Bell & J. Hooper (Eds.), Syllables and segments. New York: North Holland, 1978.

- Bell, A., & Hooper, J. Issues and evidence in syllabic phonology. In A. Bell & J. Hooper (Eds.), Syllables and segments. New York: North Holland, 1978.
- Bloomfield, L. Language. New York: Holt, 1933.
- Chomsky, N., & Halle, M. The sound pattern of English. New York: Harper and Row, 1968.
- Fant, G. Distinctive features and phonetic dimensions. In Speech sounds and features. Cambridge, Mass.: M.I.T. Press, 1973. (Originally published, 1969).
- Fischer-Jørgensen, E. Trends in phonological theory. Copenhagen: Akademisk Forlag, 1975.
- Gaitenby, J., & Mermelstein, P. Acoustic correlates of perceived prominence in unknown utterances. Haskins Laboratories Status Report on Speech Research, 1977, SR-49, 201-216.
- Greenberg, J. Is the vowel-consonant dichotomy universal? Word, 1962, 18, 73-81.
- Hjelmslev, L. The syllable as a structural unit. In Proceedings of the Third International Congress of Phonetic Sciences, 1938, 266-272.
- Hockett, C. A system of descriptive phonology. Language, 1942, 18, 3-21.
- Hockett, C. A course in modern linguistics. New York: Macmillan, 1958.
- Hooper, J. An introduction to natural generative phonology. New York: Academic Press, 1976.
- Jones, D. The phoneme. Cambridge, England: Heffer, 1950.
- Lebrun, Y. Sur la syllabe, sommet de sonorité. Phonetica, 1966, 14, 1-15.
- Lisker, L., & Abramson, A. A cross-linguistic study of voicing in initial stops. Word, 1964, 20, 384-422.
- Mermelstein, P. Automatic segmentation of speech into syllabic units. Journal of the Acoustical Society of America, 1975, 58, 880-883.
- Pike, K. Phonetics. Ann Arbor, Mich.: University of Michigan Press, 1943.
- Sievers, E. Grundzüge der Phonetik. Leipzig: Breitkopf & Härtel, 1893.
- Studdert-Kennedy, M. Speech perception. In N. Lass (Ed.), Contemporary issues in experimental phonetics (Chap. 8). New York: Academic Press, 1976.
- Trager, G., & Bloch, B. The syllabic phonemes of English. Language, 1941, 17, 223-246.
- Trager, G., & Smith, H., Jr. An outline of English structure. Studies in Linguistics: Occasional Papers 3. Norman, Ok.: Battanburg Press, 1951.
- Venneman, T. On the theory of syllabic phonology. Linguistische Berichte, 1972, 18, 1-18.
- Wedin, S., Leanderson, R., & Wedin, L. Evaluation of voice training. Folia Phoniatrica, 1978, 30, 103-112.

INFLUENCE OF VOCALIC CONTEXT ON PERCEPTION OF THE [ʃ]-[s] DISTINCTION:
III. ACOUSTIC OBSERVATIONS

Virginia A. Mann and Bruno H. Repp

Abstract. To supplement our earlier perceptual experiments (Mann & Repp, 1979; Repp & Mann, 1980), we present here additional information about the acoustic structure of our stimuli, together with further relevant measurements of natural speech and one control study. Our measurements confirm that fricative noises preceding [u] show a downward shift in spectrum relative to noises preceding [a], and that the vocalic formant transitions following [ʃ] and [s] noises differ from each other in ways that reflect the difference in place of constriction. A control experiment demonstrates that amplitude parameters (which varied a good deal in our earlier studies) have relatively little influence on perception of fricative-vowel stimuli.

A. Coarticulatory Effect of Following Vowel on Fricative Noise Spectrum

Kunisaki and Fujisaki (Note 1), Mann and Repp (1979), and Repp and Mann (1980) reported that, when listeners are presented with ambiguous fricative noises followed by [a] or [u], they perceive more instances of [s] with [u] than with [a]. Mann and Repp noted that, in the Japanese study, the perceptual context effects were considerably larger than the analogous coarticulatory effects of vocalic context on the spectral characteristics of a preceding fricative noise. To enable us to make a similar comparison for American speakers and listeners, we obtained acoustic measurements on a single speaker (LR) and reported them in Mann and Repp (1979). We present these data here once more, together with measurements on two additional speakers.

Three male native speakers of American English, all experienced phoneticians, spoke the syllables [sa], [ʃa], [sul], [ʃu] as part of a random list containing a number of other utterances. Speaker LR provided 12 tokens of each fricative-vowel syllable; speakers AA and LL, 10 each. The utterances were recorded in a soundproof booth and subsequently digitized at 10 kHz using the Haskins Laboratories PCM system. Spectral cross-sections of the fricative noises were generated in 12.8-msec steps using a Federal Scientific UA-6A spectrum analyzer with a 25.6-msec window. The frequency of the lowest prominent energy peak (presumably coincident with the first pole) was recorded for each individual section. The measurements were then averaged across all

Acknowledgment: This research was supported by NICHD Grant HD01994 and BRS Grant RR05596 to the Haskins Laboratories, and by NICHD Postdoctoral Fellowship HD05677 to Virginia Mann.

[HASKINS LABORATORIES: Status Report on Speech Research SR-62 (1980)]

sections, resulting in an estimate of the average first-pole frequency of a given noise token. For each speaker, averages and standard deviations of these estimates are displayed in Table 1.

Table 1
Acoustic variation of [ʃ] and [s] noise contingent on
the following vowel.^a

Utterance	Speaker		
	LR	AA	LL
[fa]	2405 (94)	2548 (56)	2256 (104)
[fu]	2115 (133)	2380 (88)	2028 (100)
[sa]	3773 (149)	3688 (156)	3408 (76)
[su]	3563 (116)	3124 (56)	3252 (76)

^aFrequency of lowest spectral peak (in Hz), with standard deviation in parentheses.

It is clear from inspection of Table 1 that the frequency of the first pole was consistently lower in [-u] context than in [-a] context, both for [ʃ] and for [s]. The results of t-tests showed all individual comparisons to be significant ($p < .01$), although the extent of the context effect varied somewhat across speakers. The average shifts were 229 Hz for [ʃ] and 310 Hz for [s]. By comparison, the average perceptual context effect in Experiment 1 (condition 2) of Mann and Repp (1979), expressed in terms of the first-pole dimension of the synthetic [ʃ]-[s] noise continuum employed there, was about 500 Hz. Thus, our data confirm those of Kunisaki and Fujisaki (Note 1) by showing that the perceptual effects are considerably larger than the acoustic ones, which suggests either that listeners overcompensate in perception or—more likely—that intrinsic knowledge of coarticulatory dependencies in production is not the only factor guiding perception.

B. Role of Relative Amplitude Levels

One factor that may have had some influence on our perceptual results and that we did not give sufficient attention in our earlier studies is the amplitude relationship between the acoustic components of fricative-vowel stimuli. In this section, we first describe the amplitude characteristics of the stimuli used in our earlier experiments, together with some relevant observations on natural speech, and then report a control study designed to reveal the role played by relative amplitude levels in perception of the [ʃ]-[s] distinction.

Since we did not have easy access to an automatic amplitude display, we determined relative amplitudes by excerpting a brief waveform segment (20-50 msec) from the peak amplitude region of a given stimulus portion, iterating that segment continuously, and getting a reading from the VU meter of a Crown Series 800 tape recorder. Waveform amplitudes were changed digitally to bring signals into the limited range of the VU-meter scale; an appropriate value was then added to or subtracted from the meter reading. In our earlier papers, we reported amplitude levels that had been estimated by a method that we now consider to be inaccurate. Therefore, all measurements reported here override the earlier values.

Relative amplitudes of fricative noises. For Experiment 1 in Mann and Repp (1979)--MR1 for short--and Experiment 1 in Repp and Mann (1980)--RM1 for short--the fricative noises of the nine-member synthetic continuum had identical amplitude specifications at the synthesis stage. The OVE IIIc synthesizer, however, produced output of unequal amplitude, owing to built-in dependencies between the fricative-pole amplitudes that are intended to mimic natural speech. For this reason, the peak amplitude of the noises actually increased by about 4 dB from the most [f]-like stimulus (No. 1) to the most [s]-like stimulus (No. 9). In Experiments MR2 and RM2, we decided to eliminate this amplitude gradient. In Experiment MR2, this was done by adjusting the amplitude specifications at the synthesis stage; in Experiment RM2, the digitized waveforms of the noises were adjusted instead, which had the advantage of preserving amplitude contours. In each case, later measurements indicated that we had somewhat overshot our goal, resulting in a reversed amplitude gradient (in favor of [f]) of about 2 dB. In any case, although our manipulations of the relative amplitudes of [f] and [s] might have had a slight effect on the precise location of the [f]-[s] boundary (see below), they--unlike the amplitude relationships discussed in the following paragraphs--could not have affected the magnitude of the context effect due to the periodic stimulus portion.

Relative amplitudes of periodic portions. The relative amplitudes of the two periodic portions, appropriate to [a] and [u] (or [ta] and [tu]) are an important parameter: To the extent that they had any effect on fricative perception, their effect was confounded with the effect of vowel quality. We were aware of this from the beginning but, in a well-intentioned attempt to deal with the problem, may have made it more acute. When we first compared the amplitudes of our synthetic [ta] and [tu] portions, we found [tu] to be 10 dB lower than [ta], apparently due to built-in characteristics of the OVE IIIc synthesizer. This difference seemed too large to be tolerated; therefore, we specified an amplitude for [tu] that was 10 dB higher than that for [ta]. Later measurements indicated, however, that our original assessment of relative amplitudes had been incorrect. In terms of peak amplitudes, the original [ta] and [tu] differed by only 3 dB (in favor of [ta]); hence, the modified periodic portions actually used in Experiments MR1, MR2, and RM1 differed by about 7 dB in favor of [tu].

This difference, while not disturbing to listeners, deviates markedly from the normal amplitude relationship between [a] and [u], which seems to be closely approximated by the OVE IIIc synthesizer. Representative values in the literature are 2.0-5.3 dB (Lehiste & Peterson, 1959) and 1.8 dB (Fairbanks, House, & Stevens, 1950) in favor of [a]. The relative amplitudes of

the natural [a] and [u] portions in Experiment RM2 were consistent with previous observations: The peak amplitude of [a] was, on the average, 1 dB above that of [u] for our male speaker and 3 dB above that of [u] for our female speaker. Since Experiment RM2 demonstrated a large effect of vowel quality on fricative perception, the analogous effects observed in our earlier experiments that used all-synthetic stimuli should, at least in part, have reflected the same factor. However, they may have been enhanced or reduced by the unnatural amplitude difference between the periodic stimulus portions. Below, we report a control study that examined that possibility.

Relative amplitude of fricative noise and periodic portion. The amplitudes of our fricative noises were relatively low: In Experiments MR1 and RM1, their peak amplitudes were 18-22 dB (depending on the particular noise) below the amplitude of the [ta] portion, and 25-29 dB below that of the [tu] portion. In Experiment MR2, the more [s]-like noises were further attenuated, resulting in amplitude differentials of 22-24 dB with [ta] and 29-31 dB with [tu]. In Experiment RM2, the synthetic fricative noises had a more appropriate amplitude relationship to the following (natural) periodic portions; the amplitude difference averaged 10 dB but varied from 5 to 15 dB, depending on the particular stimulus. It is possible that the relatively low noise amplitudes in Experiments MR1, MR2, and RM1 resulted in an artifactual enhancement of the vowel context effect. This issue was also addressed in the control study.

A control study. This study involved the orthogonal variation of two factors: (1) the relative amplitude of the two periodic portions, and (2) the amplitude of the fricative noise relative to the periodic portion. All amplitude modifications were performed on the digitized waveforms of synthetic stimuli used in Experiment MR1. There were four conditions, each of which contained a random sequence of 140 stimuli, resulting from the combination of seven fricative noises (stimuli 2-8 of the original nine-member continuum) with two periodic portions, presented 10 times each. In condition 1 (which replicated condition 2 of Experiment MR1) [ta] was 7 dB below [tu], and the fricative noises were 18-22 dB below [ta]. In condition 2, the [tu] portion was attenuated by 14 dB, so that it was now 7 dB below [ta]. In condition 3, all fricative noises were amplified by 12 dB. In condition 4, noise amplification was combined with attenuation of [tu]. Thus, the effect of the relative amplitudes of the periodic portions can be assessed by comparing conditions 1 and 3 with conditions 2 and 4, and the effect of the amplitude relationship between fricative noise and periodic portion can be assessed by comparing conditions 1 and 2 with conditions 3 and 4.

Eight subjects (ourselves and six colleagues at Haskins Laboratories) listened to the four tapes in counterbalanced order. Four of the listeners were familiar with the stimulus materials; the other four were relatively inexperienced. Combined across subjects, the results showed the expected overall effect of vowel context, $F(1,7) = 26.3$, $p < .001$: All subjects gave more "sh" responses in [(t)a] context than in [(t)u] context. Attenuation of [(t)u] by 14 dB resulted in a smaller vowel context effect, $F(1,7) = 10.1$, $p < .025$, primarily due to an increase in "sh" responses in the soft-[(t)u] context. However, no subject showed a complete disappearance or even reversal of the context effect in this condition. Expressed as the difference in percent "sh" responses (averaged across the seven members of the noise

continuum) between [(t)a] and [(t)u] contexts, the size of the context effect was 12.2 percent when [(t)u] was loud and 9.1 percent when [(t)u] was soft. The other experimental manipulation, a 12-dB amplification of fricative noises relative to the periodic portions, surprisingly resulted in an increase in the average context effect from 9.4 to 11.8 percent, $F(1,7) = 7.1$, $p < .05$. There was no significant interaction between the effects of the two amplitude manipulations.

Thus, the results of this control study show that our choice of amplitudes for synthetic [ta] and [tu] in Experiments MR1, MR2, and RM2 may have resulted in a slight artifactual enhancement of the context effect. However, that fact may safely be neglected, since we have now shown that context effects of nearly the same magnitude are obtained when the relative amplitudes of [ta] and [tu] are reversed. Apparently, the relatively low amplitude of our fricative noises in Experiments MR1, MR2, and RM1 was not responsible for any portion of the context effect; on the contrary, it may have reduced the effect somewhat by making [ʃ] and [s] more difficult to discriminate. Thus, the two minor artifacts may actually have canceled each other.

C. Acoustic Measurements of [ʃ] and [s] Formant Transitions

Having demonstrated large effects of formant transitions on fricative perception (Exp. RM2), we wondered about the nature and extent of the difference between the transitions characteristic of [ʃ] and [s]. As far as we know, this difference has not been described in any detail in the literature. Since [ʃ] has a more posterior place of articulation than [s], one might expect the onsets of F2 and F3 in the periodic portion to be less separate for [ʃ] than for [s], in analogy with the pattern of transitions observed for stop consonants with comparable places of articulation. To examine the extent of this difference, we first analyzed the stimuli actually used in Experiment RM2 and then, to establish the generality of our observations, measured formant onset frequencies in similar utterances pronounced by four additional speakers.

Rather than using spectrograms, which do not offer the necessary resolution, we relied on visual inspection of spectral cross-sections generated by a Federal Scientific UA-6A spectrum analyzer and displayed as point plots on a Hewlett-Packard 1300A scope, together with a computer-generated spectrogram and a waveform display. The spectral cross-sections were computed over 25.6-msec time windows, smoothed and pre-emphasized, with 12.8 msec between successive sections. The maximal resolution was approximately 40 Hz. One drawback of this equipment was that the precise location of the time windows with respect to signal onset could not be controlled. Thus, the first section included more or less of the silence preceding an utterance. Spectral peaks usually appeared only in the second section, and frequently only the third section provided a clear picture of the formant structure. Therefore, the measurements reported here are conservative with respect to any differences between utterances, since they do not derive strictly from stimulus onset but from 10-20 msec into the periodic portion.

Stimuli of Experiment RM2. The stimuli in that experiment included the periodic portions of [sa], [fa], [su], [fu] produced three times each by two speakers, one male and one female. Spectral peaks were traced through 10 consecutive cross-sections (128 msec) of each isolated periodic portion, starting with the first section that gave a clear formant pattern. (It was assumed that most formant transitions would be completed within that period; the total durations of the periodic portions varied between 300-500 msec.) The formant tracings of the three tokens of a given utterance were averaged, omitting spurious peaks not common to all three tokens. In the case of the female speaker, numerous minor peaks due to individual harmonics obscured the course of the formant transitions, especially in [a] stimuli. The formant trajectories in these stimuli were estimated by visual interpolation.

Average formant patterns are shown separately for the male and female speakers in Figure 1. Five formants could be traced for the male speaker, only three for the female speaker. The results are fairly clear. There seemed to be no systematic differences in F1 between [(f)-] and [(s)-] stimuli. F2 had a higher onset frequency in [(f)-] than in [(s)-] tokens, as predicted. F3 had a lower onset in [(f)u] than in [(s)u], which is also in agreement with the predictions: Higher F2 onset and lower F3 onset reflect a more posterior place of articulation. There was apparently no difference in F3 between [(f)a] and [(s)a], although for the male speaker F3 was so weak and inconsistent in [(s)a] that the comparison could not really be made. (Note that F3 was much more pronounced in [(f)a].) In addition, the male speaker showed a curious but consistent pattern in the higher formants of [a] stimuli: While [(s)a] exhibited only one formant, [(f)a] showed two rapidly converging spectral peaks instead. No such difference was observed in [u] stimuli, where both fricative noises were followed by similar, constant F4 and F5 resonances.

Formant onsets in a larger corpus of utterances. To confirm and extend these measurements, we recorded four additional speakers saying [sa], [fa], [su], [fu] 10 times in random order, intermixed with various other utterances. Two of the speakers (AA and LL) were experienced phoneticians; their utterances had also provided fricative noises for the measurements described in Section A. The third speaker (BHR) was one of the authors; he is a native speaker of German and pronounces [s] with a slight lisp, due to irregular dentition. The fourth speaker (SP) was a female undergraduate student; her formants (F2 and F3 only) proved to be much easier to trace than those of the female speaker in Experiment RM2. For each token of each utterance, all major spectral peaks (except F1) were recorded from the first spectral cross-section following the onset of periodicity that showed a clear pattern. (This was sometimes the first, sometimes the second.) Inconsistencies were resolved, as far as possible, by comparisons between tokens and, in some cases, by re-examination. (The measurements of AA's and LL's F2 and F3 onsets were actually done twice, with good agreement.) Means and standard deviations were calculated across the 10 tokens of each utterance, and values for [(f)-] and [(s)-] stimuli were compared by individual t-tests. The results are shown in Table 2. Significant differences ($p < .01$) between pairs of values are indicated by underlines. The results will be summarized formant by formant.

F2: [(f)-] stimuli consistently showed higher onsets of F2 than [(s)-] stimuli. In one case [(s)u] vs. [(f)u], speaker AA, five tokens showed a

MALE SPEAKER

FEMALE SPEAKER

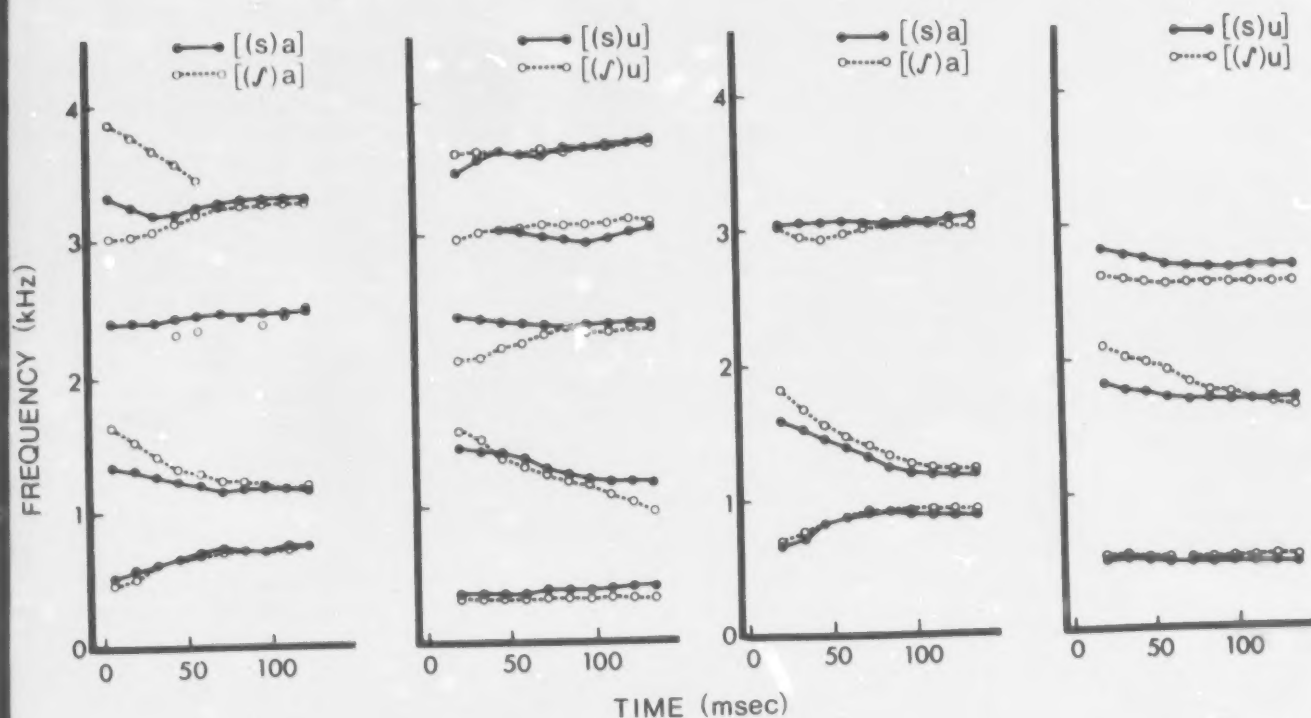


Figure 1. Formant trajectories (averaged over three tokens) during approximately the first 128 msec of the natural-speech periodic stimulus portions used in Experiment RM2.

Table 2

Average formant onset frequencies (in Hz) following [f] and [s] noises;
standard deviations in parentheses.

Utterance	F2	F3	F4	F5
<u>Speaker AA (male)</u>				
[sa]	<u>1400</u> (52) ^a	2684 (84)	[3004 (116)] ^b	<u>3772</u> (232)
[fa]	<u>1540</u> (52)	2660 (52)	3200 (52)	<u>4196</u> (112)
			{ 3032 (52) }	
[su]	1444 (72)	<u>2516</u> (36)	{ [3320 (32)] }	<u>3804</u> (60)
[fu]	{ 1740 (64) }	<u>2100</u> (108)	3016 (28)	<u>4000</u> (60)
	{ 1436 (76) }			
<u>Speaker LL (male)</u>				
[sa]	<u>1400</u> (60)	<u>2528</u> (52)	<u>3644</u> (84)	4264 (192)
[fa]	<u>1708</u> (52)	<u>2208</u> (80)	<u>2948</u> (52)	4256 (52)
[su]	1448 (52)	<u>2524</u> (44)	<u>3440</u> (60)	4216 (108)
[fu]	<u>1764</u> (72)	<u>2144</u> (32)	<u>2956</u> (56)	{ 4248 (108) }
				{ [3624 (20)] }
<u>Speaker BHR (male)</u>				
[sa]	<u>1388</u> (56)	<u>2732</u> (56)	3232 (92)	4000 (100)
[fa]	<u>1472</u> (28)	{ <u>2872</u> (80) }	3232 (52)	[3896 (232)]
		{ [2192 (132)] }		
[su]	1612 (48)	2520 (28)	<u>3208</u> (40)	3780 (52)
[fu]	1684 (68)	2452 (72)	<u>3072</u> (80)	{ [3344 (72)] }
				{ [3996 (72)] }
<u>Speaker SP (female)</u>				
[sa]	<u>1720</u> (80)	2872 (80)		
[fa]	<u>2044</u> (44)	2924 (40)		
[su]	<u>2076</u> (60)	<u>3028</u> (64)		
[fu]	<u>2236</u> (64)	<u>2860</u> (72)		

^aUnderlined values are significantly different from each other ([f] vs. [s]) by a t-test, $p < .01$.

^bValues in brackets indicate spurious or unreliable peaks.

striking difference, the other five did not, suggesting a rather abrupt transition; therefore, two averages are reported.

F3: F3 had consistently lower onsets in [(f)u] than in [(s)u]. The relationship between [(f)a] and [(s)a] was more variable, just as Figure 1 had suggested. One speaker (LL) showed a much lower onset in [(f)a] than in [(s)a]. Two speakers (AA and SP) showed no clear difference. Speaker BHR showed two peaks in [(f)a] instead of one (usually both at the same time), one strikingly lower, the other somewhat higher than the single peak in [(s)a].

F4: One speaker (LL) differed from the other two males by showing extremely large differences in the F4 region: F4 onsets for [(f)-] were 500-700 Hz lower than for [(s)-]. Speaker BHR showed a similar, smaller difference in [u] stimuli only. Speaker AA showed a variable pattern: a weak and inconsistent F4 in [(s)a] but not in [(f)a], and an extra peak in [(s)u] but not in [(f)u].

F5: One speaker (AA) showed higher onsets of F5 for [(f)-] than for [(s)-], despite large variability in [(s)a] tokens. Speaker LL, on the other hand, showed no such differences at all. (Spurious peaks at a lower frequency were noted in [(f)u].) Speaker BHR showed weak and inconsistent F5 peaks in [(f)-] utterances, at two different locations in the case of [(f)u], so that no clear picture emerges.

In summary, [f] transitions are characterized by a higher F2 onset, a lower F3 onset (at least in [u]), and, in some speakers, by a lower F4 onset and/or a higher F5 onset than [s] transitions. This pattern, at least as far as the lower formants are concerned, is consistent with the fact that [f] has a more posterior place of articulation than [s]; it confirms our earlier observations depicted in Figure 1.

REFERENCE NOTE

1. Kunisaki, O., & Fujisaki, H. On the influence of context upon perception of voiceless fricative consonants. Annual Bulletin of the Research Institute for Logopedics and Phoniatrics, 1977, 11, 85-91.

REFERENCES

- Fairbanks, G., House, A. S., & Stevens, E. L. An experimental study of vowel intensities. Journal of the Acoustical Society of America, 1950, 22, 457-459.
- Lehiste, I., & Peterson, R. E. Vowel amplitude and phonemic stress in American English. Journal of the Acoustical Society of America, 1959, 31, 428-435.
- Mann, V. A., & Repp, B. H. Influence of vocalic context on perception of the [f]-[s] distinction: I. Temporal factors. Haskins Laboratories Status Report on Speech Research, 1979, SR-59/60, 49-64.
- Repp, B. H., & Mann, V. A. Influence of vocalic context on perception of the [f]-[s] distinction: II. Spectral factors. Haskins Laboratories Status Report on Speech Research, 1980, SR-61, 65-84.

THE ROLE OF AUDITORY MEMORY IN SPEECH PERCEPTION AND DISCRIMINATION*

Robert G. Crowder†

Abstract. Theoretical statements on the role of auditory short-term memory in speech perception and discrimination are reviewed. The significant empirical phenomena that fall within the domain of this review are categorical perception, the differences among phonetic classes in auditory persistence, contextual effects in phonetic labeling, and selective adaptation. Many of these observations can be organized by a theory of auditory representation in memory that draws on principles of recurrent lateral inhibition.

INTRODUCTION

The operations for testing memory and perception in an experimental context are generally the same even though we often are smug enough to think we understand the differences between the two. In each case the person is asked to make some response on the basis of what he thought occurred in the past. The problem is very nicely illustrated in the discrimination procedure traditionally used for establishing the phenomenon of categorical perception (Liberman, Harris, Hoffman, & Griffith, 1957): The punch line in that demonstration, of course, is that people's discrimination of tokens from a synthetic /b, d, g/ continuum is no better than would be predicted on the basis of their phonetic labeling of those tokens. Items from within a phonetic category cannot be discriminated, although physically equally close items that happen to cross a category boundary can.

The Dual Coding Model

It is when we examine the discrimination task used in the categorical perception experiment that we see the inseparability of perception and memory. The technique of choice for years was the ABX discrimination task, where the subject must decide whether a probe stimulus at the end of the presentation triad resembles more the first or the second in the triad. Formally, this procedure is a degenerate case of testing memory for temporal position. In a less degenerate case, the subject might be read lists of eight letters and be

*This paper was presented at the International Symposium on the Cognitive Representation of Speech, Edinburgh, July, 1979.

†Also at Yale University, New Haven CT.

Acknowledgment. The preparation of this paper and the research in it were supported by NSF Grant BNS77-07062 to R. G. Crowder, and by NICHD Grant HD-01944 and BRS Grant RR-05596 to Haskins Laboratories. The author thanks Bruno Repp, Alice Healy, and Virginia Walters for their comments and assistance.

asked in what position he heard a probe letter repeated at the end of the list -- an A-B-C-D-E-F-G-H-X procedure. One might complain that the analogy is ridiculous because an eight-letter memory load is qualitatively different from a two-letter memory load. However, once it is realized that not only is memory for letters (or phonetic labels) involved but also memory for sounds, the characterization of the ABX task as a memory paradigm becomes quite serious.

This insight was the contribution of Fujisaki and Kawashima in a set of brilliant technical reports in 1969, 1970, and 1971 (Notes 1, 2, and 3). Their model for the ABX task is shown in Figure 1 taken from Fujisaki and Kawashima (Note 1). A moment's study reveals that the model proposes a serial consultation of phonetic, then auditory memories during the task. The subject first decides whether the two initial stimuli (A & B) are different phonetic segments. If so, he classifies the probe X phonetically and compares its label with those stored for A and B. If A and B were given the same phonetic label, then the subject must rely on auditory short-term memory to decide which resembles X more.

As these ideas are by now quite well known, I shall not pause to describe the independent evidence in their favor. Much of it is due to the experimental work of Pisoni (1971, 1973, 1975). The clear theoretical advance in Pisoni's work was to relate the "short-term memory" model of Figure 1 to the difference between stop consonants and vowels in categorical perception. (Fry, Abramson, Eimas, and Liberman, 1962, showed that unlike stop consonants, steady-state vowels can be discriminated well in the ABX paradigm even if they come from the same phonetic category.)

Vowels and Consonants in Auditory Memory

The notion advanced by Pisoni was that if speech discrimination depends on auditory and phonetic short-term memory as proposed by Fujisaki and Kawashima, then the categorical perception result could be derived from a process model of discrimination. In particular, obtaining categorical perception will depend on whether there is auditory short-term memory that survives until the comparisons of A, B, and X are made. If we assume that auditory memory for consonants is either absent or severely limited, it follows that they should be discriminated no better than they can be differentially labeled. If, on the other hand, vowels benefit from better or longer representation in auditory memory, then the subject should be able to supplement phonetic memory with auditory memory when A and B are from the same phonetic category.

Support from A-X (same/different) discrimination. This logic led to the prediction (Pisoni, 1973) that the time separating the to-be-discriminated sounds should affect performance. Pisoni switched to the simpler, A-X (same/different) discrimination task and varied the time interval between the two tokens from zero to two seconds. The stimuli were steady-state vowels or stop consonants. He found that the time separation of the two items had a large effect on within-category discrimination accuracy for vowels, with dramatic losses in performance between the quarter-second and two-second intervals. However, within-category performance on the consonants was relatively unaffected by delay and uniformly poor. Between-category

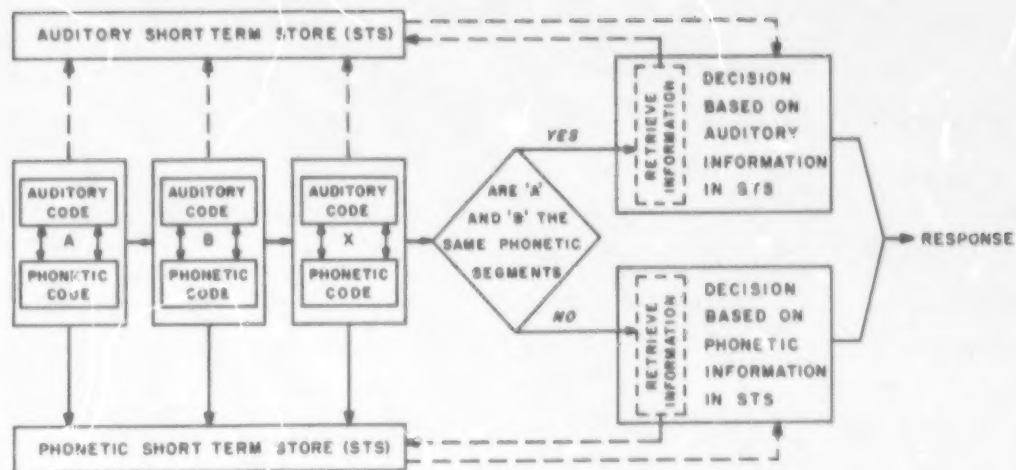


Figure 1. The Fujisaki-Kawashima process model for auditory and phonetic memory in A-X discrimination.

discrimination for the stops was quite high and not systematically affected by delay. These results make the case that it is the memory difference between vowels and consonants that underlies their classic differentiation in the categorical perception experiment.

Evidence from immediate memory. Independent support for the memory difference between vowels and stops came from some immediate-memory experiments I reported a few years ago (Crowder, 1971, 1973a, 1973b). Morton and I (Crowder & Morton, 1969) had identified two experimental techniques for evaluating auditory memory in the context of list-memory experiments in which subjects had to repeat series of memory-span items. We argued that the advantage of auditory over visual presentation, an advantage confined to the last positions within the list, could be understood as the consequence of a brief auditory short-term memory store not available in the visual-presentation condition. We showed that the addition of a redundant item, called a stimulus suffix, seemed to remove, or mask, this auditory advantage.

In one experiment (Crowder, 1971) I showed that the modality and suffix effects were absent when people have to remember lists of items distinguished only by stops (BAH BAH GAH DAH BAH DAH GAH) but that these effects showed up normally when comparable lists contained vowel distinctions (BEE BIH BIH BOO BEE BEE BOO). Subsequent experiments (Crowder, 1973a, 1973b) showed the following:

1) The result with stops does not depend on using initial stops because the outcome was the same with VC syllables (AHB, AHD, AHG,...).

2) Lists based on fricative contrasts give an intermediate result between the sizeable suffix and modality effects for steady-state vowels and the absence of these effects for stops.

3) The degree of auditory memory for vowels can be influenced by the particular tokens used. Long-duration (300 millisecond) vowels seem to allow better auditory-memory representation than short-duration vowels (50 milliseconds). This result matches findings by Fujisaki and Kawashima and Pisoni that vowels behave more like stops when their duration is sharply cut back.

My original interpretation of these results was in terms of a special speech processor (see also Liberman, Mattingly, & Turvey, 1972) that was selectively recruited for stop consonants; the idea was that the stop-vowel differentiation in memory could be understood in linguistic, rather than psychoacoustic, terms.

The discriminability hypothesis. Darwin and Baddeley (1974) argued, on the contrary, that my results were better understood on the psychacoustic hypothesis of a universal auditory memory subject to degradation with the passage of time. They showed that other types of consonants besides stops did benefit from auditory memory representation in the suffix experiment, if they were drawn from a memory set with highly discriminable phonetic classes (including nasals and fricatives). Secondly, they showed that if vowels were selected from nearby locations in vowel space, so as to make them less easily discriminable than vowels from a wide set of locations, the auditory memory

representation was reduced. Finally, they showed that diphthongs gave ample evidence of auditory memory even though they are based on transient, rather than steady-state, information.

Darwin and Baddeley (1974) concluded that auditory memory holds a crude representation of the signal, whatever speech sound it is, for a short period during which it is subject to increasing degradation. As they put it,

The experiments reported here suggest that the result of this degradation is in some way to blur the information held in acoustic memory. After some degradation has taken place there may be sufficient information left to distinguish between a number of very different items, but perhaps not enough information left to distinguish between the same number of more similar ones... It may be that the different estimates of the duration of acoustic memory, that previous workers have found, are due to the different auditory resolution that their tasks required. (p. 54)

The psychoacoustic, degraded tape-recording, model is certainly the most parsimonious explanation for these results at present. My own thinking has come around to Darwin and Baddeley's on this point. In a recent theoretical paper (Crowder, 1978) I have made explicit that the form of an auditory memory is like a crude sound spectrogram or a crude Fourier analysis. This position will be explained in more detail below. For now, I want to continue with the line of reasoning based on Fujisaki and Kawashima's model because we now have reason to believe it is wrong.

Predictability and Context Effects in Labeling

A shortcoming of the Fujisaki-Kawashima-Pisoni position is that for any discrimination task, the contribution of auditory memory is assessed by measuring the difference between the discrimination performance predicted from phonetic labeling and the discrimination performance that is obtained. In other words, the extent to which performance exceeds a phonetic "categorical perception" baseline is the measure of how much auditory memory contribution there is. As Macmillan, Kaplan, and Creelman (1977) and others have commented, this is a dangerously circular way to define auditory memory; it is a sort of performance wastebasket.

Repp, Healy, and Crowder (1979) undertook to manipulate the availability of auditory-memory information for A-X vowel discrimination directly. They noted that Pisoni had shown that increasing the delay between the two stimuli to be discriminated reduces discrimination; this is presumably because of decay in the auditory short-term memory (since phonetic memory should not be affected by times on the order of a second). Also, Pisoni showed that interposition of a distractor sound between the AB pair and the X probe degrades discrimination performance, again, presumably through a masking of auditory memory not unlike the stimulus suffix effect. What Pisoni had not shown was that operations designed to affect the availability of auditory memory also affect the degree of fit between discrimination predicted from phonetic labeling and obtained discrimination.

Say that a sufficient delay and masking were introduced into the A-X task for vowels that no auditory memory at all remained for A at the time X occurred. We should then be able to predict exactly how successful the subject's "same-different" discrimination would be by knowing the probability that he assigned same or different phonetic labels to the two stimuli (which is measured separately). This prediction follows directly from the logic of Figure 1.

As Repp et al. (1979) put it,

If accuracy of A-X discrimination can be predicted from phonetic labeling (identification), provided that both are better than chance, we may conclude that perception is categorical. The interesting possibility is that although discrimination shows a surplus over identification when auditory memory is present, vowel perception will be categorical when auditory memory has been removed. (p. 131)

In their first experiment, Repp et al. tested discrimination of steady-state vowels from the /i, I, ε/ continuum under four main conditions. The delay between the two A-X stimuli was about one-half second in two of the conditions and about two seconds in the other two conditions. Orthogonally, the delay interval was either left silent or was filled with a potentially interfering masking sound, /y/.

Both manipulations, increased delay and interference, reliably impaired discrimination performance. Furthermore, in the condition with both interference and the long delay, obtained discrimination was almost depressed to the level of discrimination predicted from a purely phonetic or "categorical perception" model. Repp et al. were not satisfied with this result at face value, however, for two reasons. First, the stimuli they used had not been chosen at equal spacings along the /i, I, ε/ continuum (for reasons that need not detain us here). Secondly, the data on phonetic labeling for their stimuli were taken from the conventional one-at-a-time or "out of context" identification test. Since vowel identification is known to be influenced by context (Eimas, 1963, for example), the fairer test would be to measure labeling (phonetic identification) of the sounds under exactly the circumstances the labels might be used in discrimination--that is, in the context of the A-X discrimination tapes. In other words, the Fujisaki-Kawashima model assumes a covert phonetic process during discrimination; we wanted our overt phonetic process to be coming from the same discrimination context.

In the second experiment, there were only two main conditions, one with a short unfilled interval and another with a long filled interval; the delays and interference conditions were similar to those of the first study. But in this study there were two phonetic identification tasks from which to predict discrimination, one for the short unfilled interval and one for the long filled one.

As in the first experiment, the results showed that discrimination far exceeded predictions from the conventional (out-of-context) identification

test in the short, unfilled condition, where there was presumed to be abundant auditory memory information. Discrimination in the long, filled condition was much worse, as in the earlier study; however, although discrimination approached single-item identification predictions here, we could not say performance was categorical by this criterion.

The main outcome concerned the identification data taken from the A-X discrimination tapes. In this measurement, subjects were asked to listen to both sounds in the pair and then respond with phonetic labels for each of the two stimuli. We were startled by the large, contrastive context effects that were evident in these data in the condition with a short unfilled interval between the two stimuli. In this condition, for example, the probability of labeling a stimulus as the sound /ɛ/ in the interior of the 13-item continuum was .55 if the second stimulus of the pair came from the opposite (/i/) side of the continuum. The probability dropped to a value of .27 when the second member of the pair was a more extreme example of /ɛ/.

The retroactive and proactive effects were not statistically distinguishable although the retroactive effects were numerically stronger. (Retroactive contrast is when the second of the items affects identification of the first item and proactive contrast is the reverse.) These contrast effects were much stronger in the short, unfilled condition than in the long, filled condition (although not completely absent in the latter). The fact that we got larger contrast effects at the short interval than at the long interval is important for it suggests a basis of the contrast effects in auditory memory. It will be proposed below that contrast should be expected to the degree that two vowels occupy auditory memory together.

The second main result of the Repp et al. experiment concerned the relation between obtained discrimination performance and predictions for discrimination based on "in context" labeling, where subjects listened to the same A-X discrimination tapes but tried to label each of the two sounds phonetically, rather than respond "same" or "different." We took these identification data and then scored them as if subjects had been doing discrimination--if the two stimuli were assigned the same label we counted it as a "same" response, otherwise "different." The finding was that these in-context identification data predicted obtained discrimination very well but not perfectly. The discrepancy between predicted and obtained discrimination was the same in the two delay conditions. Of course, since different phonetic identification was occurring in the two delay conditions (because of differential context effects), the predictability was measured separately for the two conditions. For stimuli from the mid-range of the thirteen-stimulus continuum, predictions were almost perfect--that is, subjects instructed to discriminate on the basis of physical identity did no better than when instructed to give the two items phonetic labels. At the extremes there were minor departures from categorical perception, particularly at the /ɛ/ end of the scale.

CONCLUSIONS

Conclusions about categorical perception. Whether one chooses to decide from these results that vowels are perceived categorically depends on which of two definitions of categorical perception one decides to embrace. By the traditional definition--whether discrimination on a physical criterion can be predicted from labeling on a phonetic criterion--vowels are perceived just about as categorically as consonants are. (There is quite typically a "discrimination surplus" in studies of stop consonants, too, although the surplus is not large, as it was not here.) Repp et al. observed that if one defines categorical as absolute, in the dictionary sense of the word, then vowels are not perceived categorically, since they are subject to such powerful context effects.

Conclusions about auditory memory in speech perception. Now it remains to carry these results back to our consideration of auditory memory and its role in speech perception. We think the Repp et al. results are decisive against the Fujisaki-Kawashima-Pisoni model: According to that model, discrimination is based on consultation of two short-term memory codes, first a phonetic code and then an auditory code. In our two interference conditions (short-unfilled versus long-filled) we demonstrably manipulated the integrity of the auditory memory. Therefore we should have found a much larger "surplus" of discrimination over identification in one condition than in the other. Instead, we found the discrepancies between discrimination and identification to be equal in the two conditions.

Repp et al. could not reach a firm conclusion about the basis for discrimination performance in their experiment; various mixtures of auditory and phonetic processing were proposed as alternatives. One possibility merits attention here because it is particularly novel and clear: The entire contribution of auditory memory may reside in context effects influencing application of phonetic labels and the actual discrimination may be entirely phonetic.

Why, then, do subjects bother to derive phonetic readings for the sounds in a physical A-X discrimination task? Repp et al. suggested that there might be an inherent priority for linguistic levels of analysis when stimuli include the potential for such analysis. The Stroop effect would be an example of such priority for linguistic analysis even when it is counterproductive. The modest discrepancy between the obtained discrimination and discrimination predicted on phonetic labeling just shows that this priority is not mandatory for all subjects on all trials.

DISCUSSION

Relation to Selective Adaptation

The hypothesis that auditory memory has a primary function in exerting contextual effects upon phonetic labeling is consistent with results that have been interpreted as evidence for selective adaptation of neural feature detectors in speech perception (Eimas & Corbit, 1973). We begin this argument

with the comment that the formal operations are the same: In a selective adaptation experiment, the subject is played many repetitions of a token from one extreme of some continuum, say from the /ba/ end of the /ba-pa/ VOT continuum. The finding is that after such an experience, the /ba-pa/ boundary is shifted towards the adapting (/ba/) stimulus. In other words, an ambiguous item from near the boundary sounds more like /pa/ if the subject has been listening to unambiguous instances of /ba/. The Repp et al. result shows the same thing--a sound from the /i-I/ boundary sounds less like the former if it is preceded immediately by a token from the left end of this continuum than otherwise.

There are two main differences between the selective adaptation paradigm and the phonetic contrast paradigm. First, the contrast effects most typically occur with vowels whereas the adaptation experiments typically use stops. However, we can dismiss this difference right away because Eimas (1963) showed that stops are subject to qualitatively the same contrast effects from neighboring identification targets as vowels are. The second difference is that the adaptation experiments use repeated presentation of the adaptation (context) stimulus whereas the contrast demonstrations are mainly interested in one prior or subsequent context item.

Selective adaptation and contrast. A recent article by Diehl, Elman, and McCusker (1978) suggests that the use of repeated adaptors may largely be a matter of superstition or faith in the fatigue metaphor. In their first experiment, these authors showed that a /b-/d/ boundary shift could be obtained with a single adaptor separated by 1.5 seconds from the test item. They also showed that the effect depended on using extreme tokens as adapting stimuli rather than more ambiguous tokens, as is true in the selective adaptation literature. Diehl et al. showed in their second experiment that the "cross-series-adaptation" result could be simulated in the single-item context paradigm. Finally, their third experiment showed that substantial contrast effects still occurred when the "adapting" stimulus occurred after the test stimulus. This result clearly is anomalous from the Eimas and Corbit (1973) feature detector perspective, but as we shall see below it is consistent with the contrast view. (Recall that Repp et al. found comparable forward and backward contrast effects in their experiment on steady-state vowels.)

Anchors, contrast, and selective adaptation. Another recent publication, by Sawusch and Nusbaum (1979), adds weight to the contrastive interpretation of the selective adaptation result. These authors were interested in "anchor effects" and their possible relation to contrast. They used stimuli from the /i, I, ε / continuum and presented either a balanced series for identification or a series in which a good exemplar of one of the three phonetic categories occurred in an eight-to-one ratio to all the other 12 stimuli. There were four seconds between items. The finding was that boundaries shifted consistently toward the category chosen as anchor and accordingly over-represented in the test series. That is, in an identification series "heavy" with tokens of /i/, items from the /i/-/I/ boundary would tend to be called /I/ more than otherwise.

Sawusch and Nusbaum showed that these effects were symmetrical for the three different anchoring stimuli, /i, I, ε / and that the effect did not depend on the number of response categories available. In a final study, they

instructed subjects that there would be many extra presentations of the anchoring vowel; however, this intended manipulation of response criterion had no effect on the boundary shift. The authors favor an interpretation based on the idea that a previous stimulus in auditory memory provides a "ground" against which new, ambiguous stimuli are heard in contrast. The next section proposes a model that makes such an assumption explicit.

A Model for Auditory Memory and Contrast

I have recently developed a theory of backward masking in auditory memory taking off from experiments using a list-memory technique (Crowder, 1978). Without reviewing that line of research here, I want to illustrate the major features of the theory in the event it might be useful in the present context. The major assumption is that auditory events are laid out on a two-dimensional array that encodes their time of arrival and the physical channel on which they came. The slightly novel feature of the assumption is that there might be a memory representation for time that is somehow neurally spatial. The simple-minded diagram of Figure 2 shows the hypothetical representation. Unfortunately it has not been possible for me to speak precisely about the second organizational dimension for this memory representation, called "Physical Channel" in Figure 2. I mean to capture with this dimension the traditional selective-attention sense of communication channel--the dimension on which two voices of the same sex are moderately discrepant, two voices of opposite sexes are more discrepant, and on which a speech signal and a noise are extremely discrepant. One might look to the experiments of Treisman and her collaborators in the 1960's (for example, Treisman & Geffen, 1967) for functional information on what distinctions can be called channel distinctions.

The main process assumption that accompanies this representation assumption is borrowed wholesale from the retina of the Horseshoe Crab (see Cornsweet, 1970). I assumed that the memory representations defined by their locations on the time-by-channel array behaved as do units in the two-dimensional retinal array in the visual system, namely in accordance with the laws of recurrent lateral inhibition.

This accounts trivially for backward masking among auditory memory representations; if two activations are close enough together in time and similar or identical in channel of arrival, they will mutually inhibit one another. There are other facts in the list-memory setting which are nicely accommodated by the model, as can be verified in Crowder (1978). But my purpose now is to return with this model to problems addressed earlier in this paper.

Auditory memory and different speech sets. For example, there is the evidence that auditory memory is differentially rich for different sets of speech sounds. It will be recalled that this was a cornerstone of the Pisoni amplification of Fujisaki and Kawashima's model, that vowels are somehow better represented in auditory memory than stops. Independently, Crowder, and Darwin and Baddeley turned in evidence directly relating auditory memory to "discriminability" of memory sets.

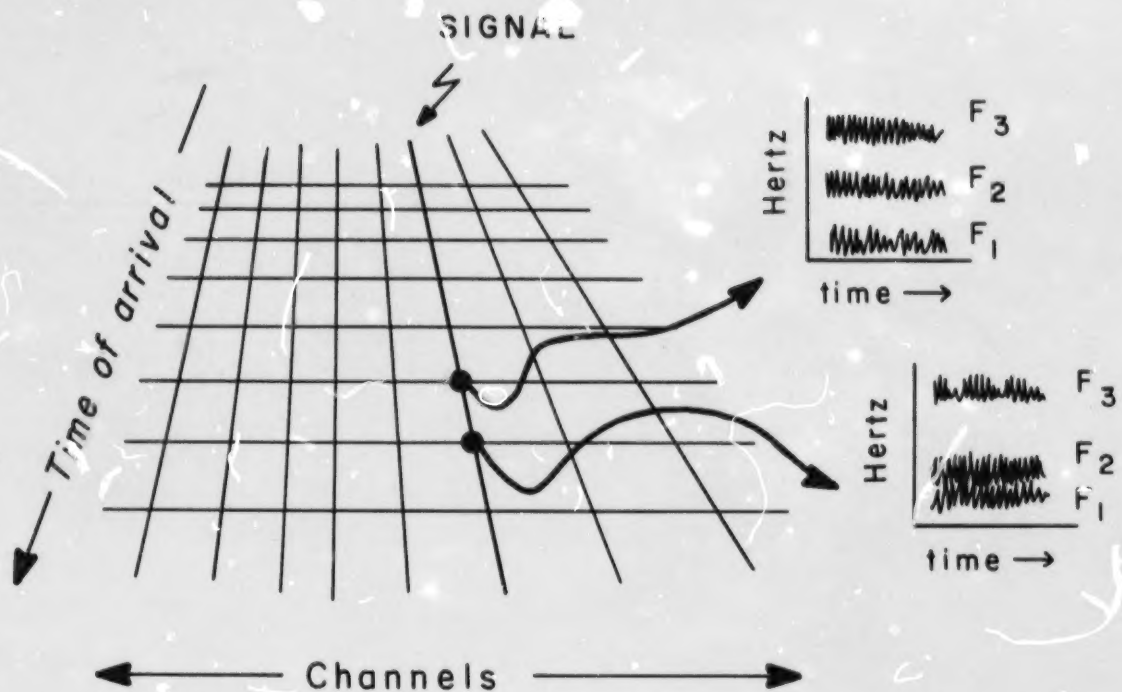


Figure 2. A hypothetical representation for auditory memories in two-dimensional neural space. Entries are classified by channel of entry and time of arrival.

In the model of Figure 2, I made the assumption that the contents of the representations stored at intersections of the grid are some form of crude spectral analysis, perhaps comparable to a smudged sound spectrogram. This assumption was partly necessary because, in a masking situation, the level of activation of representational units by itself is not an informative event (unlike the state of affairs on the retinal mosaic, where degree of activation, rate of firing, is the important commodity.) It does little good to know that a particular channel was active at a particular moment, we need to know what happened there. Thus, the memory representation entered on the grid had to be something spectral. But it had to be only crudely accurate because a high fidelity representation would be too good: Remember the system of auditory memory has to be useful for highly distinct steady-state vowels but relatively useless for stop consonant distinctions. The best candidate seemed to be something like a sound spectrogram with a bit of perturbation--the sort of representation that would preserve long-term relations between formant frequencies, as can be used to cue steady-state vowels, but not subtle or transient formant changes, as are the cues for the less discriminable speech distinctions.

Contrast effects. As of this writing, there is one missing process assumption: It seems natural to assert that the lateral inhibitory influence of two representations that are close enough on the layout of Figure 2 should produce contrast in phonetic labeling. That is, if a representation for a good solid /i/ is close to a representation for a borderline /i/ or /I/, the inhibitory influence of the first on the second might result in contrast. However, exactly why has not been clearly worked out. Only one rather wild conjecture has occurred to me:

The inhibition might be frequency specific in that memory representations that are close to one another on the grid inhibit spectral components in common only. It would be as if one took spectrograms of the two mutually inhibiting sounds and held them up together. Where they were identical in frequency, the intensity of both would decline. If the two sounds in question came from the same source channel, most of the energy would tend to be in overlapping frequency bands, hence the main backward masking result. With our two vowels /i/ and the ambiguous /i/-/I/, however, although the high overlap in spectral energy would tend to weaken both, there would be specific changes from which we could derive contrast.

In this vowel continuum, it is F1 that is changing most informatively between /i/ and /I/. Remember that the memory representation is a smudged spectrogram, which is the equivalent of a high bandwidth formant structure. Thus the prototype /i/ and the borderline /i/-/I/ have considerable overlap in F1, though the latter will have an overall higher F1. The critical assumption is that the two stimuli cancel each other out to some extent in their common F1 "territory" and the high F1 region of the latter stimulus is left relatively intact. Thus, the surviving spectral information about the ambiguous token will, by virtue of the frequency-specific lateral inhibition, show a higher mean F1 and will accordingly be more confidently called /I/ than when there was no contrasting sound in auditory memory.

Independent support for this conjecture must be sought; however, it does all fit together. The fact that Repp et al. observed symmetric forward and

backward contrast effects is consistent with the general spirit of Figure 2, in which time has been transformed into a neurally-spatial dimension. The fact that selective adaptation and phonetic contrast effects are both larger when the adapting (context) stimulus is an acoustically "strong" contrast to the test stimuli than when it is only a "weak" contrast (see Diehl et al., 1978) is to be expected (although the application of these ideas to VOT has not yet been worked out). If the spectral overlap between adaptor and test stimulus were too high, there would be little information "escaping" lateral inhibition and forming the basis for contrast. That is, if our prototype /i/ were too close to the boundary itself, there would be little spectral information in the ambiguous /i/-/I/ to form the basis for a functional elevation of the F1 in the latter.

In my judgment, the neural feature-detector interpretation of selective adaptation has been severely vexed by the finding (Ades, 1977) that adaptation with one voice did not produce a boundary shift in a test stimulus presented by another voice. If it were a true phonetic feature detector in question, it should fire and get fatigued for, say, voiced stops, whoever happens to be saying them. On the other hand, the theory being sketched here anticipates this finding. If the two stimuli in question are separated in source channel, on Figure 2, we should expect no lateral inhibition and no contrast effect.

Further Evidence for the Model

Although Pisoni and others have shown that increasing the time separation between two stimuli in the A-X discrimination paradigm leads to poorer performance, I made the conjecture above that A-X performance (in the Repp et al. study) might be based on phonetic and not auditory information. The Pisoni result would then have to be attributed to the following chain of events: The delay between A and X affects the amount of auditory contrast between them and hence the amount by which they tend to be given contrastive labels (if they are truly different). These phonetic labels are essentially serving to sharpen differences when the two are close together in auditory memory. These sharpened differences (like edge sharpening in vision) affect phonetic labeling, which is then subsequently used in responding "same" or "different."

By this account, it would be advisable to have some information on labeling behavior and its dependence on the interval between A and X when labeling is collected from A-X discrimination tapes. A recent unpublished experiment of mine provides such information.

Auditory memory in phonetic labeling. In a recent experiment I played subjects pairs of steady-state vowels from a continuum ranging from /a/ to /ae/, which most subjects felt comfortable labeling with the vowel sounds from cot, cut and cat. In one half of the experiment I had them performing A-X discrimination; we shall not consider those results here. In another condition, I asked them to apply a phonetic label to the second member of each pair only. If one considers the 13 vowels on the continuum to be numbered from 1 (cot) to 13 (cat), the two items in a pair were always in ascending order. That is, the subject would hear stimuli paired 1-3, and 8-11, for example, but never 3-1, or 11-8. The items chosen for pairs were the 13

possible "same" pairs, all the 11 possible two-step discriminations, and all of the 10 possible three-step discriminations.

The stimuli were 300 milliseconds long and the main variable was the length of the silent interstimulus interval between A and X, which varied from .2 to 4.7 seconds in .5 second steps. The question was whether, on trials where the two stimuli were different, the boundaries would be shifted towards the first member of the pair and whether this influence would be stronger when the items of the pair were close together in time.

The labeling functions in Figure 3 show simply that subjects heard the 13 stimuli in three categories, as I had intended; these data are taken only from the trials in which the two stimuli were identical, the "same" trials. Notice that one category boundary occurs between items 6 and 7 and the other between items 9 and 10.

For purposes of examining boundary shifts, I collapsed the two boundaries as if there were only one, yielding an artificial average boundary between 7 and 8, for the data of Figure 3. For each interstimulus-interval condition, the combined boundary was estimated by linear regression using items 3 through 6 for the first boundary and items 8 through 11 for the second boundary. The ten interstimulus intervals were further collapsed into five to reduce variability.

The results are shown in Figure 4 which gives the combined boundary value separately for the "same" and "different" trials as a function of delay interval. As expected, the proximity of a prior context stimulus made no difference when it was identical to the subsequent, test stimulus. However, this was not the case when the two stimuli were different. At the shorter intervals, the boundary was shifted towards the left, that is towards the prior stimulus, in accordance with the contrast data under consideration here. At the longer interstimulus intervals, after about three seconds, the contrast disappeared and labeling was not different in the "same" and the "different" conditions.

The disappearance of contrast effects at about three seconds agrees perfectly with other evidence I have obtained with these and other stimuli, that a decay in same-different discrimination performance caused by increasing interstimulus interval is asymptotic at about three seconds.

A Final Comment

Is it self-deception, or do the conclusions reached here make matters of theory a bit simpler than we thought before? Apparently, a theory of auditory memory developed within the context of verbal memory experiments can be carried, without gross distortion, into the speech-discrimination paradigm. In that new context, the theory permits a unified framework for considering categorical perception, the dual-coding hypothesis, contrast in phonetic identification, selective adaptation, and anchor effects. This theoretical framework is not yet clear of the rocks: One fundamental unsolved problem is whether the contrast effects of the sort observed in Figure 4, or in Repp et al. (1979), have an auditory or a phonetic basis. What if the person were thinking of the /i/ sound when he heard an ambiguous /i-I/ token? Another

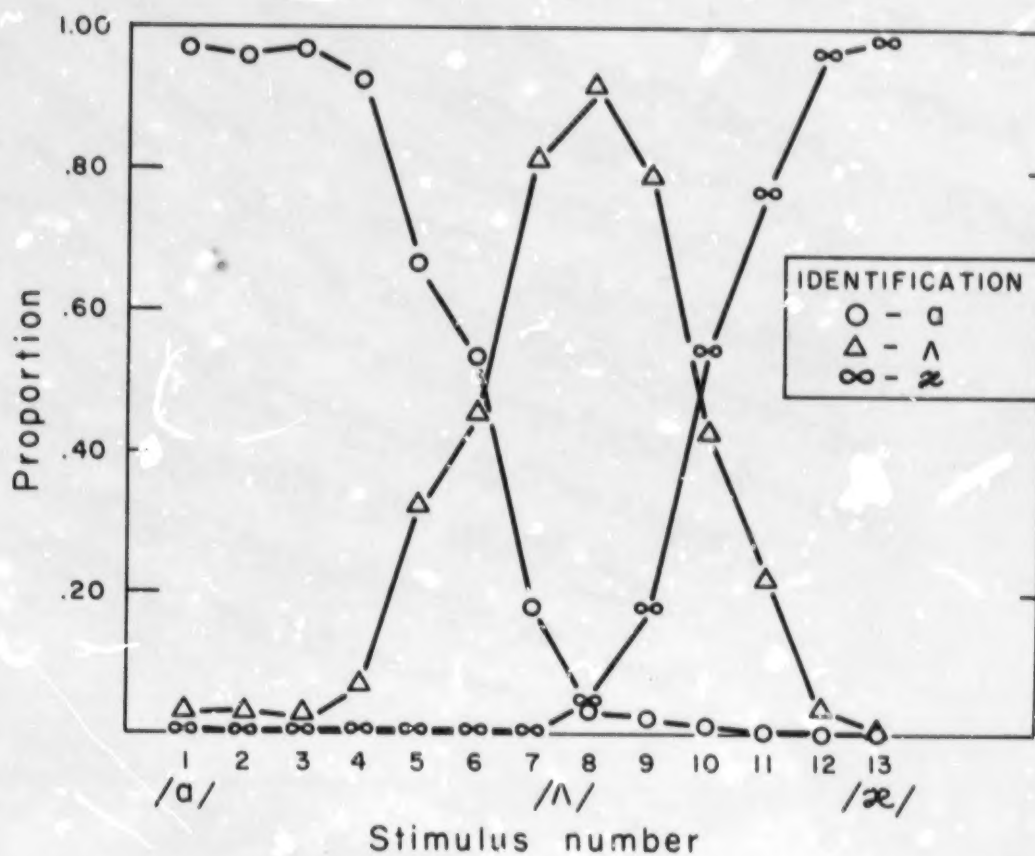


Figure 3. Phonetic labeling on "same" trials, collapsed over interstimulus intervals.

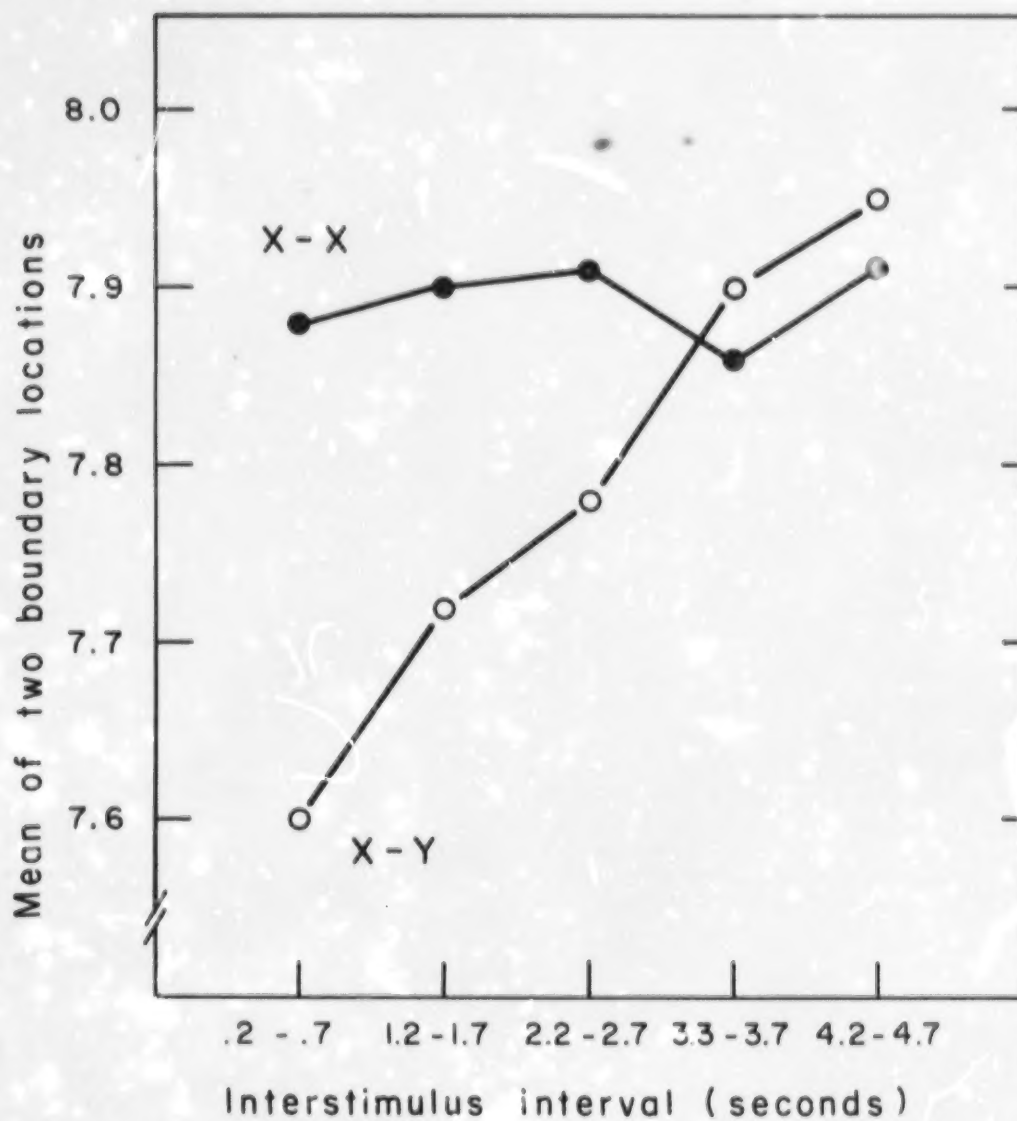


Figure 4. Collapsed boundary values for "same" (X-X) and "different" (X-Y) trials when subjects were labeling the second of two vowels separated by varying interstimulus intervals.

problem is how to apply the contrast model based on lateral inhibition to the VOT continuum. Without that application we might as well give it up right at the beginning.

It is clear, in any case, that considerations of auditory memory are here to stay in speech perception research.

REFERENCE NOTES

1. Fujisaki, H., & Kawashima, T. On the modes and mechanisms of speech perception. Annual Report of the Engineering Research Institute, University of Tokyo, 1969, 28, 67-73.

2. Fujisaki, H., & Kawashima, T. Some experiments on speech perception and a model for the perceptual mechanism. Annual Report of the Engineering Research Institute, University of Tokyo, 1970, 29, 207-214.

3. Fujisaki, H., & Kawashima, T. A model of the mechanisms for speech perception -- Quantitative analysis of categorical effects in discrimination. Annual Report of the Engineering Research Institute, 1971, University of Tokyo, 3, 59-68.

REFERENCES

- Ades, A. E. Vowels, consonants, speech, and nonspeech. Psychological Review, 1977, 84, 524-530.
- Cornsweet, T. N. Visual perception. New York: Academic Press, 1970.
- Crowder, R. G. The sound of vowels and consonants in immediate memory. Journal of Verbal Learning and Verbal Behavior, 1971, 10, 587-596.
- Crowder, R. G. The sounds of speech in immediate memory. Journal of Experimental Psychology, 1973, 98, 14-24. (a)
- Crowder, R. G. Precategorical acoustic storage for vowels of short and long duration. Perception and Psychophysics, 1973, 13, 502-506. (b)
- Crowder, R. G. Mechanisms of auditory backward masking in the stimulus suffix effect. Psychological Review, 1978, 85, 502-524.
- Crowder, R. G., & Morton, J. Precategorical acoustic storage (PAS). Perception and Psychophysics, 1969, 5, 365-373.
- Darwin, C. J., & Baddeley, A. D. Acoustic memory and the perception of speech. Cognitive Psychology, 1974, 6, 41-60.
- Diehl, R. L., Elman, J. L., & McCusker, S. B. Contrast effects in stop consonant identification. Journal of Experimental Psychology: Human Perception and Performance, 1978, 4, 599-609.
- Eimas, P. D. The relation between identification and discrimination along speech and nonspeech continua. Language and Speech, 1963, 6, 206-217.
- Eimas, P. D., & Corbit, J. D. Selective adaptation of linguistic feature detectors. Cognitive Psychology, 1973, 4, 99-109.
- Fry, D. B., Abramson, A. S., Eimas, P. D. & Liberman, A. M. The

- identification and discrimination of synthetic vowels. Language and Speech, 1962, 5, 171-189.
- Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. The discrimination of speech sounds within and across phoneme boundaries. Journal of Experimental Psychology, 1957, 54, 358-368.
- Liberman, A. M., Mattingly, I. G., & Turvey, M. T. Language codes and memory codes. In A. W. Melton & E. Martin (Eds.), Coding processes in human memory. Washington, D.C.: Winston, 1972.
- Macmillan, N. A., Kaplan, H. L., & Creelman, C. D. The psychophysics of categorical perception. Psychological Review, 1977, 84, 452-471.
- Pisoni, D. B. On the nature of the categorical perception of speech sounds. Unpublished doctoral dissertation. University of Michigan, 1971.
- Pisoni, D. B. Auditory and phonetic memory codes in the discrimination of consonants and vowels. Perception and Psychophysics, 1973, 13, 253-260.
- Pisoni, D. B. Auditory short-term memory and vowel perception. Memory & Cognition, 1975, 3, 7-18.
- Repp, B. H., Healy, A. F., & Crowder, R. G. Categories and context in the perception of isolated steady-state vowels. Journal of Experimental Psychology: Human Perception and Performance, 1979, 5, 129-145.
- Sawusch, J. R., & Nusbaum, H. C. Contextual effects in vowel perception I: Anchor induced contrast effects. Perception and Psychophysics, 1979, 25, 292-302.
- Treisman, A., & Geffen, G. Selective attention: Perception or response? Quarterly Journal of Experimental Psychology, 1967, 19, 1-17.

TALKER-INDEPENDENT INFORMATION FOR VOWEL IDENTITY*

Robert R. Verbrugge⁺ and Brad Rakerd⁺

Abstract. A male and a female talker each produced a set of /bVb/ syllables (V=/i,I,e,ε,æ,a,ʌ,ɔ,ʊ,u/) at metronome-controlled rates. The central 60% of the voiced regions of these syllables was edited out, producing single-talker silent-center syllables (cf. Strange, Jenkins, & Edman, 1978). "Hybrid" silent-center syllables were then created by appending the final 20% of one talker's syllable to the initial 20% of the other talker's production of the same syllable, again with a 60% silent region intervening. Vowel identification errors were relatively low for both the single-talker (22.4%) and the hybrid (27.0%) silent-center syllables, though errors in each case exceeded those found for the original (unedited) syllables (9.3%). Hybrids were generally heard as coherent syllables from a single source. The error pattern for the hybrids could not be treated as a combined function of errors on the initial portions (55.4% on average) and the final portions (72.9%) treated as independent elements. Results indicate that considerable information for vowels is contained in the dynamic regions early and late in a syllable. Moreover, they suggest that this information does not simply subserve computation of an underlying normalized target; the information may specify properties of production dynamics that are invariant for a vowel for all talkers of a dialect. To the extent that such information is available, normalization in perception is unnecessary.

The primary acoustic information for English vowels is typically defined in terms of a set of formant frequency values measured at the center of a syllable. Each vowel is characterized, in essence, as a fixed articulatory posture or cavity shape with characteristic resonant frequencies. The perceiver, in turn, is typically viewed as a system that measures formant frequency values and then categorizes a vowel following computations on those measures.

There are several lines of evidence, however, that suggest that these cross-sectional measures do not exhaust the primary acoustic information used by English listeners. Listeners seem to make considerable use of time-varying, or "dynamic," information in syllable structure. For example,

*This paper was presented at the 99th meeting of the Acoustical Society of America, Atlanta, Georgia, 21-25 April, 1980.

⁺Also University of Connecticut, Storrs.

Acknowledgment: This research was supported by NICHD grant HD-01994 and BRS grant RR-05596 to Haskins Laboratories.

listeners often show more accurate identification for vowels produced in consonantal environments than for vowels produced in isolation (e.g., Strange, Verbrugge, Shankweiler, & Edman, 1976; Strange, Edman, & Jenkins, 1979). A steady-state formant pattern, with no dynamics, carries relatively little information for vowel identity; steady-state vowels show poorer identification than comparable isolated vowels produced dynamically (Lehiste & Meltzer, 1973; Shankweiler, Verbrugge, & Studdert-Kennedy, 1978).

The most dramatic evidence for the importance of dynamic information was obtained by Strange, Jenkins, and Edman (1978), in a study of silent-center syllables. These syllables were produced by substituting silence for the central 50-65% of the voiced regions of /bVb/ syllables--in effect, eliminating the entire quasi-steady-state region typically thought to contain the essence of the vowel. The results were remarkable. Listeners spontaneously integrated the initial and final portions across the silent gap; they typically heard a single utterance, though with an interruption. In addition, vowel identification for the silent-center syllables was remarkably accurate, approaching performance on the unedited syllables. This demonstrates that the initial and final transitional regions of a syllable are alone sufficient to support vowel perception to a high degree, and that the syllable centers are expendable.

Two possible explanations for this finding provide the motivation for the present study. First, it is possible that the listener uses the transitional regions of a syllable to extrapolate to the formant frequencies that were the intended targets of the talker. Lindblom (1963), for example, has suggested that formant trajectories can be treated as exponential functions that decelerate toward (or accelerate from) asymptotic target frequencies. By hypothesis, these asymptotes could be computed by the listener if the formants fail to reach them due to rapid speech or destressing (cf. Lindblom & Studdert-Kennedy, 1967). In short, according to this target-extraction hypothesis, the only informational role of the formant trajectories is to aid the extraction of an underlying set of spectral targets.

A second possible explanation is that time-varying spectral variables are part of the essential definition of a vowel, not simply a heuristic for getting to a vowel defined in steady-state terms. Vowels can be viewed as articulatory events, as gestures that manifest a characteristic organization of forces over the articulators. Thus, the vowels of a dialect may be distinguished by different styles of movement, as well as by the positions attained as a result of movement. These styles of movement will produce characteristic changes in acoustic variables, and these styles of acoustic change, in turn, may provide substantial information to the perceiver about a vowel's identity. The silent-center results are compatible with this event-perception approach to vowel perception; in fact, the results suggest that time-varying information alone is sufficient to support accurate identification.

To test the competing claims of the target-extraction hypothesis and the event-perception hypothesis, we constructed "hybrid" silent-center syllables by pairing the initial and final portions from syllables spoken by two different talkers, a man and a woman. We will first describe the experimental materials and then return to the predictions each approach would make.

A man and a woman each recorded multiple tokens of eleven /bVb/ syllables containing the vowels /i, I, e, ε, æ, a, ɔ, ʊ, u/. Their productions were controlled in rate by a metronome. For each vowel, we selected a pair of syllables, one from each talker, that were closely matched in duration. Figure 1 displays a waveform of one of the unedited syllables, the woman's /bæb/. Following the general strategy of Strange et al. (1978), we defined the initial and final portions of the syllable in the following way. The initial portion included the release burst of the initial /b/, plus the first 20% of the voiced region of the syllable. The final portion included the final 20% of the voiced region, plus the closure interval and release burst of the final /b/. In the silent-center syllables, the central 60% of the voiced region was edited out.

There were a total of five experimental conditions, as illustrated in Figure 2. The oscillogram at the top is an example of a whole syllable--the original, unedited production. Second from the top is the corresponding silent-center syllable; in effect, this pairs the initial portion and final portion of the same talker's production, at the original temporal interval. Next is a hybrid silent-center, formed from the female initial portion and the corresponding final portion of the male's syllable. The temporal interval was the same as that for the single-talker silent-center syllable. All possible female-male and male-female combinations of this kind were included in the hybrid condition. The isolated initial portions and final portions comprised the fourth and fifth conditions.

In each condition, a separate group of listeners heard a randomized sequence of the tokens. They were asked to identify a single vowel that best characterized each token they heard. They did so by circling one of eleven alternative /bVb/ words written in English orthography. Listeners in the edited syllable conditions were told that the tokens consisted of pieces of real speech productions.

We return now to the predictions one might make regarding perception of the hybrid syllables. According to the target-extraction hypothesis, a hybrid syllable should be very disruptive perceptually. Because the male and female have different target values, the initial portion and final portion will point to different targets. This is illustrated in Figure 3. On the left are spectrograms of the male and female productions of /bæb/. The patterns at the right illustrate their differences in formant values by pairing the left half of the male's spectrogram with the right half of the female's, and vice versa. (In the hybrid syllables, of course, the central 60% of these pairings was silenced, contrary to what is pictured in Figure 3.) It is clear that the target formant values for the two talkers are very different. Spectral measures indicate that, on average, the female's formant frequencies were 17% higher than the male's. In the hybrid syllables, the target values extracted from the initial and final portions would differ, on average, by that amount, and it is unlikely that the two portions could be integrated perceptually. Thus, the target-extraction hypothesis predicts that separate sources will be perceived and that vowel identification will be based on perception of the two portions as independent elements.

The event-perception hypothesis suggests an intriguing alternative. A pair of talkers who speak a common dialect may produce a vowel with a common

style of change, in spite of differences in vocal tract size and articulator mass. As a result, they may show a common patterning of change in acoustic variables, even though their spectral ranges differ. If so, it is conceivable that a hybrid syllable will preserve that abstract patterning of change, even though the formant frequency range shifts abruptly between the initial and final portions. Thus, the event-perception hypothesis, in this strong form, predicts that the two portions will be integrated perceptually, and that accuracy of vowel identification will be high, perhaps as high as for the single-talker silent-centers.

The overall results for the five listening conditions are displayed in Figure 4. Each bar denotes the mean percent errors in vowel identification for the indicated condition. Errors averaged 9.3% for the whole syllables and 22.4% for the silent-centers. These results replicate those of Strange et al. (1978), who also found a comparatively high level of accuracy on silent-centers.

Errors on the hybrid silent-centers averaged only 27.0%; this was not significantly different from the 22.4% error rate for the single-talker silent-centers. Thus, whatever information is preserved in the single-talker syllables seems to be preserved in the hybrids; accuracy is essentially unaffected by the change in source. This suggests that some characteristic talker-independent information for vowel identity has been preserved in these constructions. It is also interesting to note that these chimeral hybrids are typically perceived as utterances by a single person--usually a male--and more particularly, a male who is prone to abrupt changes in pitch register. The information for a unitary articulation is apparently strong enough to define a unitary source, overcoming other evidence that might indicate a change in source.

The data for the initial and final portions, summarized on the right hand side of Figure 4, provide an important control and validation for these conclusions. Errors in each case were very high, averaging 55.4% on the initial portions, and 72.9% on the final portions. This indicates, at the outset, that acoustic change per se is not sufficient to guarantee high identification, nor is it sufficient to provide a listener with formant transitions that point toward or away from certain target values. The two portions provide substantial information only when both are present, as in the silent-centers.

One might argue, of course, that the silent-centers offer a listener two chances at identifying the vowel, and that this accounts for the higher level of accuracy. If we assume that judgments on the two portions are independent (as we are led to do by the target-extraction hypothesis), we can predict from our results on the initial and final portions that errors on the silent centers and hybrids would average 64%; this is much higher than the 22% and 27% error rates that we observed.¹ It is clear that the two portions are not independently perceived; listeners relate the structure of the initial portion to that of the final portion in a unitary percept.

In general, these results pose a challenge to any approach that takes the extraction and normalization of target values as the central enterprise of vowel perception. The hybrid syllables specify discrepant targets, yet they

cohere perceptually and are identified with relatively high accuracy. It could be argued, in defense of the normalization view, that each transition serves to normalize the target it specifies, and that the two normalized targets would therefore not be discrepant. However, if the initial transition bears rich information about one vocal tract source, and the final transition bears rich information about a different source, it is difficult to understand why the two sounds are integrated at all, let alone why they are heard as an utterance by a single talker.

Several possibilities follow from this analysis. First, if vowels are defined predominantly by a style of articulatory change that is invariant for all speakers of a dialect, then a normalization step in perception is unnecessary. A perceptual system designed for measuring time-varying acoustic variables, rather than center formant frequencies, could directly register the vowel information conveyed by any talker.

Second, the transient regions of a syllable may provide acoustic information that is specific to the vowel, without necessarily being information about a target state. A rough analogy may be drawn to the role of onset transients in the identification of musical instruments. The dynamic structure of these transients carries more information about instrument identity than does the steady-state region of a sustained tone (Grey & Gordon, 1978; Luce & Clark, 1967; Saldanha & Corso, 1964). More to the point, the transients do not simply aid the extraction of steady-state timbre; the information they provide is different in kind. In the case of vowels, we can similarly argue that the time-varying structure of a talker's onset transients is specific to the vowel being spoken and does not simply subserve identification based on steady-state formant structure.

Finally, these results suggest that the perception of source continuity cannot be attributed simply to the detection of continuities in formant structure and pitch contour (cf. Dorman, Cutting, & Raphael, 1975; Darwin & Bethell-Fox, 1977). The perception of source continuity may also be influenced by a patterning over acoustic variables that specifies a continuity of dialect, rate, or prosody. These types of constraint on speech patterning are not marked acoustically by simple continuities of pitch and formant structure, but are realized complexly by pitch, formant structure, and many other acoustic variables. Acoustic evidence for a violation of constraints on dialect, rate, or prosody could underlie perception of a discontinuity in source. In this study, the productions of the male and female talkers were closely matched for each of these types of constraint. Preservation of that common structure in the hybrids was apparently sufficient to specify a single source.

While our results to date are preliminary, they are a striking demonstration of the importance of dynamic information for vowel identity (and perhaps also for source identity). Considerable research is needed to pinpoint the time-varying information that listeners use in each case.

REFERENCES

- Darwin, C. J., & Bethell-Fox, C. E. Pitch continuity and speech source attribution. Journal of Experimental Psychology: Human Perception and Performance, 1977, 3, 665-672.
- Dorman, M. F., Cutting, J. E., & Raphael, L. J. Perception of temporal order in vowel sequences with and without formant transitions. Journal of Experimental Psychology: Human Perception and Performance, 1975, 104, 121-129.
- Grey, J. M., & Gordon, J. W. Perceptual effects of spectral modifications on musical timbres. Journal of the Acoustical Society of America, 1978, 63, 1493-1500.
- Lehiste, I., & Meltzer, D. Vowel and speaker identification in natural and synthetic speech. Language and Speech, 1973, 16, 356-364.
- Lindblom, B. E. F. Spectrographic study of vowel reduction. Journal of the Acoustical Society of America, 1963, 35, 1773-1781.
- Lindblom, B. E. F., & Studdert-Kennedy, M. On the role of formant-transitions in vowel recognition. Journal of the Acoustical Society of America, 1967, 42, 830-843.
- Luce, D., & Clark, M., Jr. Physical correlates of brass-instrument tones. Journal of the Acoustical Society of America, 1967, 42, 1232-1243.
- Saldanha, E. L., & Corso, J. F. Timbre cues and the identification of musical instruments. Journal of the Acoustical Society of America, 1964, 36, 2021-2026.
- Shankweiler, D., Verbrugge, R. R., & Studdert-Kennedy, M. Insufficiency of the target for vowel perception. Haskins Laboratories Status Report on Speech Research, 1978, SR-55/56, 103-111.
- Strange, W., Edman, T. R., & Jenkins, J. J. Acoustic and phonological factors in vowel identification. Journal of Experimental Psychology: Human Perception and Performance, 1979, 5, 643-656.
- Strange, W., Jenkins, J. J., & Edman, T. R. Dynamic information specifies vowel identity. Journal of the Acoustical Society of America, 1978, 63, S5. (Abstract)
- Strange, W., Verbrugge, R. R., Shankweiler, D. P., & Edman, T. R. Consonant environment specifies vowel identity. Journal of the Acoustical Society of America, 1976, 60, 213-224.

FOOTNOTES

¹These estimates assume that independent judgments are made on each portion and that the final response choice is made at random between the initial and final alternatives. If the wise homunculus always based its choice on the more informative portion (the initial portion), predicted errors on the hybrids would still be 55%. Lower estimates can be derived for models (unlike the target-extraction view) that predict non-independence of judgments based on the two portions. Such models would have to argue why the two portions would be perceptually contingent, given the formant and pitch discontinuities that should shear them apart (cf. Dorman et al., 1975; Darwin & Bethell-Fox, 1977). The event-perception hypothesis provides a strong explanation for perceptual integration, and argues that responses are based on a singular judgment.

FIGURE CAPTIONS

Figure 1. Waveform of an unedited syllable, the woman's /bʌb/. Initial and final portions were produced by editing the syllables at the nearest zero-crossing to the points illustrated here. In silent-center syllables, the center 60% of the voiced region was replaced by silence.

Figure 2. Sample tokens from each experimental condition. (The waveforms derive from the male and female utterances of /bʌb/.)

Figure 3. Spectrograms of male and female productions of /bʌb/ are presented on the left. To create the patterns at the right, the spectrograms were cut at the center of the voiced regions; initial and final halves were interchanged to create male-female and female-male juxtapositions. These patterns illustrate the disparity between the target formant levels for the two talkers.

Figure 4. Mean percent errors in vowel identification for the five listening conditions. Errors are pooled over all eleven vowels and over the tokens from both talkers.

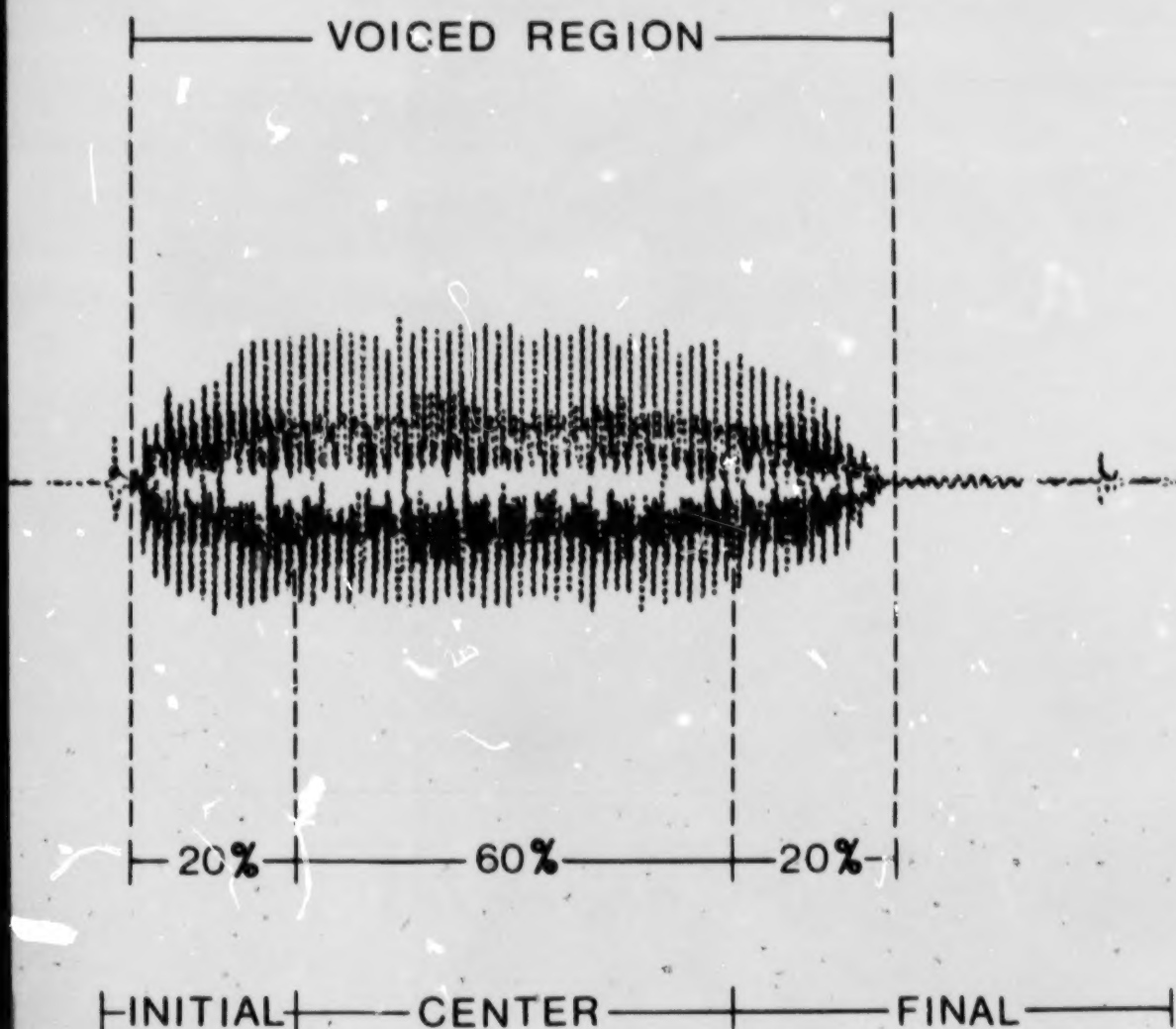


Figure 1

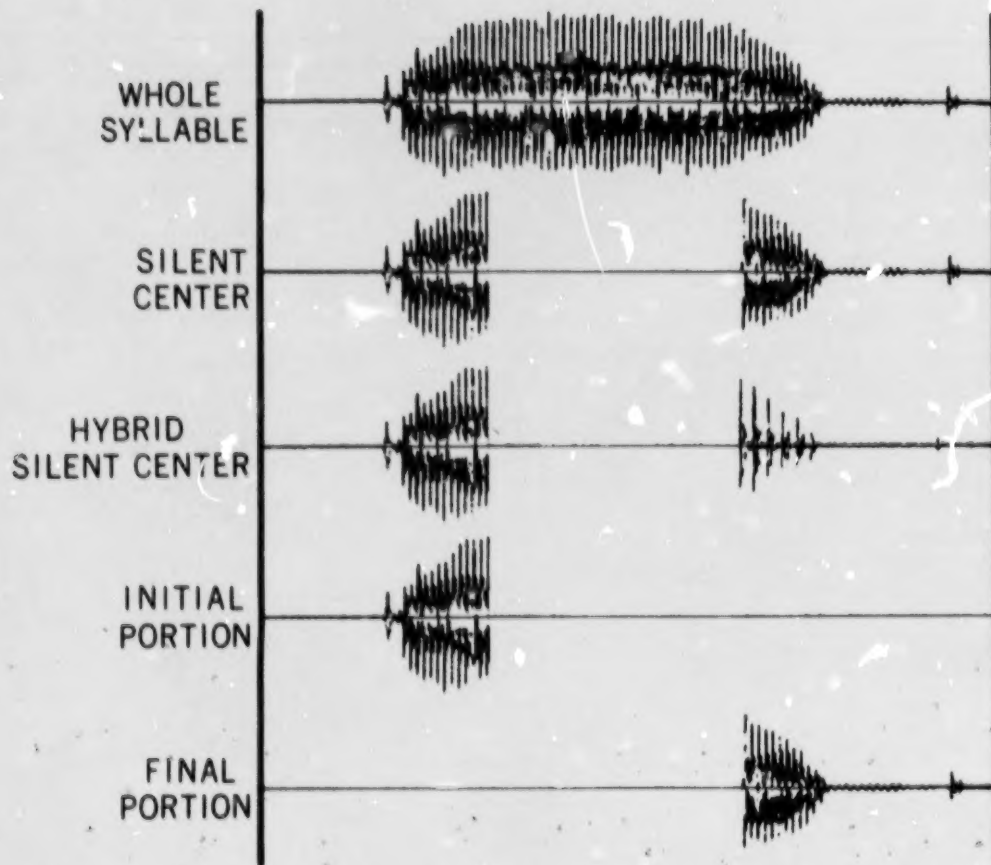


Figure 2

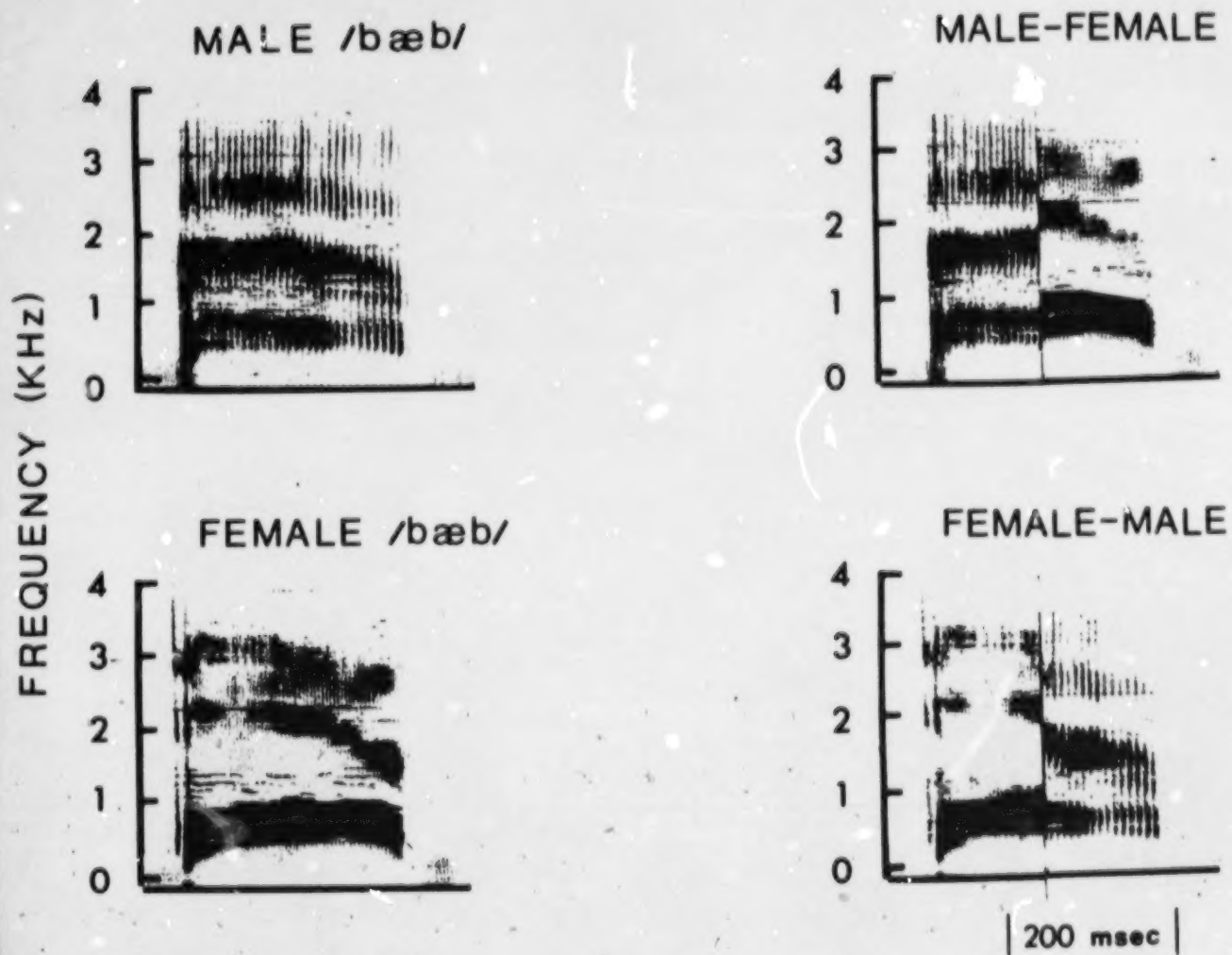
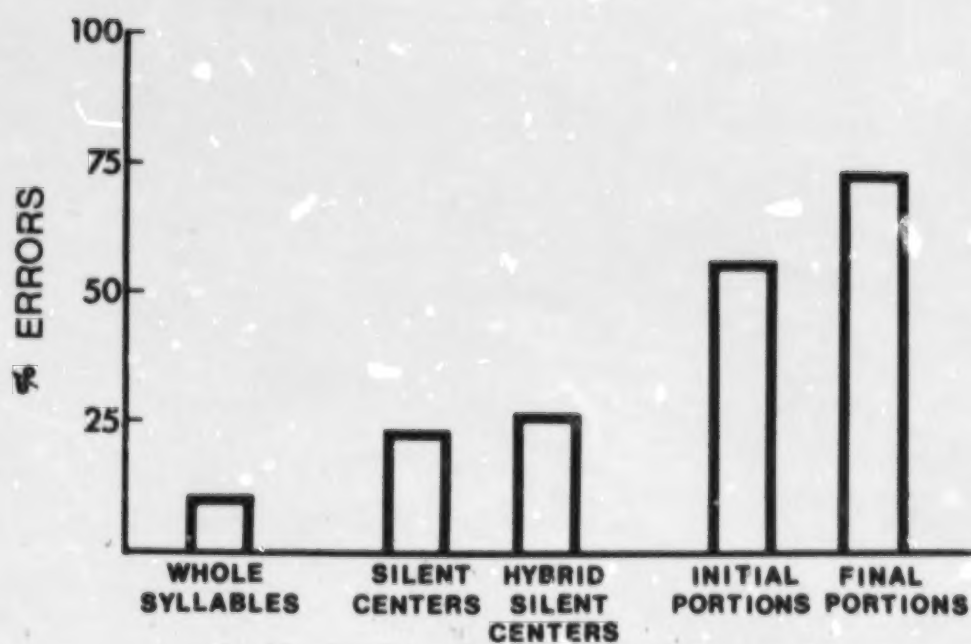


Figure 3



217

Figure 4

CONTEXT-CONDITIONED SPECIFICATION OF VOWEL IDENTITY*

Robert R. Verbrugge,+ Donald P. Shankweiler,+ and Carol A. Fowler++

Abstract. Consonantal environment does not aid vowel identification universally, but preserves and sharpens the information for vowels differentially. In a vowel monitoring task, listeners heard one of two test series: /pVp/ syllables or isolated vowels. During each block of test trials, listeners monitored for one of the nine vowel types, checking "Yes" when they recognized an instance and "No" otherwise. For both misses and false alarms, there was a strong interaction between vowel type (close vs. open) and context type (/pVp/ vs. /V/). The open vowel pairs (/ɛ/-/æ/, /ɔ/-/ɑ/) showed consistently superior recognition in the /pVp/ consonantal environment; others did not. The results suggest that constraints on coarticulation may determine the precision with which vowels can be identified.

Over the past several years, we and our colleagues have reported a variety of studies on the sources of information for vowel identity (Verbrugge, Strange, Shankweiler, & Eiman, 1976; Strange, Verbrugge, Shankweiler, & Edman, 1976; Fowler & Shankweiler, 1978; Shankweiler, Verbrugge, & Studdert-Kennedy, 1978; Strange, Jenkins, & Edman, 1978). As other papers in this session attest (Pisoni, Carrell, & Simnick, 1979; Macchi, 1979), our research has engendered some controversy. It will be helpful to begin by summarizing what the central goal of this research has been. Our goal has been to demonstrate that listeners use dynamic information when perceiving vowels. We have conceived of vowels, in articulatory terms, as gestures, defined predominantly by styles of movement rather than by fixed configurations. We have similarly argued, in the case of perception, that listeners attend to dynamic information--by which we mean that listeners detect acoustic information for articulatory movement and timing, as well as information for articulatory position. Acoustic "targets" can provide partial information for vowel identity, but they are not sufficient for distinguishing vowels as gestures (cf. Strange et al., 1978).

In order to assess the role of coarticulatory dynamics in vowel perception, we have compared the identification of vowels in consonantal environments with vowels produced in isolation. Using a multiple-choice identification task, we have found that vowel identification tends to be more accurate and less variable in a consonantal environment than in isolation (Strange et

*This paper was presented at the 97th meeting of the Acoustical Society of America, Cambridge, Mass., June 11-15, 1979.

+Also University of Connecticut, Storrs.

++Also Dartmouth College, Hanover, New Hampshire.

Acknowledgment. This research was supported by NICHD grant HD-01994 and BRS grant RR-05596 to Haskins Laboratories.

al., 1976). One possible explanation for this result is that coarticulation of consonant and vowel gestures leads to a more precise specification of the dynamic properties that distinguish vowels. Several possible sources for such an effect can be identified. For example, the production of a vowel might be more tightly constrained when its timing is coordinated with that of a consonant, and when common articulators are employed in the execution of each gesture. As a result, the vowel would be executed with less variability and acoustic information for it would be more precise. (There is some evidence that acoustic properties of isolated vowels are more variable than those of vowels in consonantal context; see Strange, Edman, & Jenkins, 1979; Tiffany, 1959.) In addition, if the consonant and vowel interact in nonlinear ways, the nature of the coupling between them could provide further information about each participant. This suggests, in turn, that the perceptual influence of a consonantal context need not be constant in type or degree, but will vary systematically with specific pairings of consonants and vowels.

The study we summarize here is a first step toward understanding the coarticulatory contingencies. Its focus is on the perception of vowels in a symmetric labial-stop environment. Rather than using a multiple-choice identification task (as we have done in the past), we designed a monitoring task for this study. Our intent was to eliminate the possibility that orthographic confusions could confound the results, and, if anything, to bias listeners' conditions in favor of the isolated vowels.

Listeners in the monitoring task heard one of two test series: /pVp/ syllables or isolated vowels. The test items were natural productions from a single male talker, who produced five tokens of each of nine vowels. Each test block contained ten repetitions of each of the nine vowels, for a total of 90 randomized trials. During each test block, listeners monitored for instances of one of the nine vowel types, checking "Yes" when they recognized an instance and "No" otherwise. Before each test block they were informed which vowel they were to listen for; the instructions included three examples of English words that contain the vowel, followed by three productions of the vowel in isolation. Listeners heard a total of nine blocks of 90 trials, monitoring in each block for only one vowel.

Listeners can make two types of error in a task of this kind. A miss is a failure to recognize an instance of the vowel being monitored (i.e., saying "No" when you should say "Yes"). A false alarm is an erroneous acceptance of a vowel as an instance of the vowel being monitored (saying "Yes" when you should say "No").

The left-hand side of Figure 1 shows average percent misses for all nine vowels for the two context conditions in the monitoring task. Misses averaged 11% for the isolated vowels and 9% for the vowels in /pVp/ environment. (False alarm scores, which are not presented in Figure 1, showed a small overall difference in the same direction.) The right-hand side of Figure 1 presents the average percent errors for an identification task using the same materials (from the data of Strange et al., 1976); errors in that study averaged 33% for the isolated vowels and 10% for the vowels in consonantal context. In each task, monitoring and identification, an overall advantage was observed for vowels in /pVp/ context. However, there was a dramatic difference between the two tasks in errors on the isolated vowels; these were

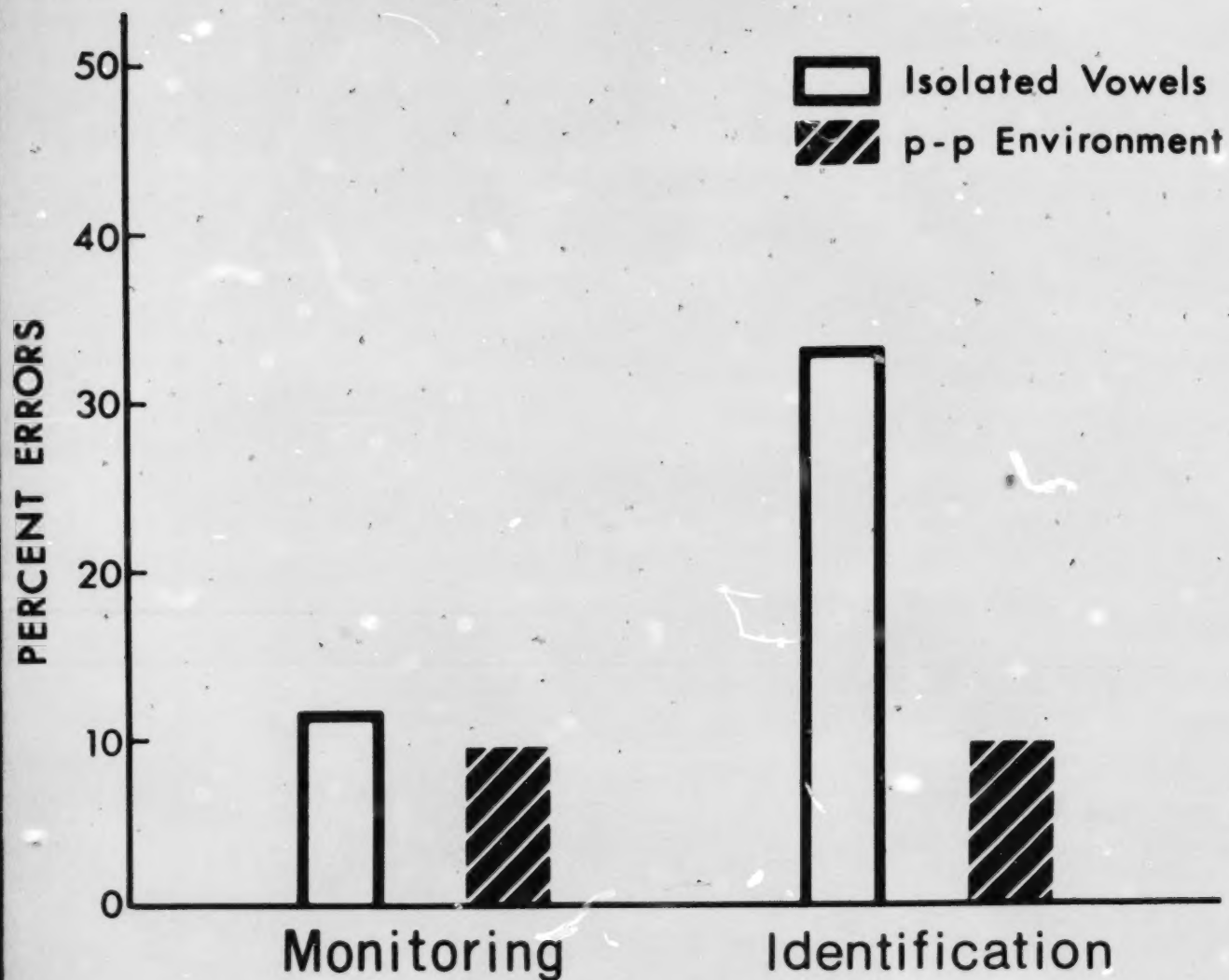


Figure 1. Mean percent errors on isolated vowels and vowels produced in /pVp/ environment. Results are depicted for a monitoring task (misses) and an identification task.

sharply reduced in the monitoring task, with the result that the overall difference between the two context conditions was almost completely neutralized.

Several task variables need to be considered as possible contributors to this change in overall errors for the isolated vowels. For example, there may have been a response-sheet bias favoring /pVp/ syllables in the identification tasks; or, conceivably, orthographic labeling problems may have affected isolated vowels differentially in the identification task. It is not immediately clear which task variables contributed most strongly to the difference in error scores. Certainly, it is not clear why problems in representing the vowels orthographically would have been so much more severe for the isolated vowels than for the /pVp/ syllables on the identification task.

The profound effect that task variables can have on error scores suggests caution in accepting any single task as a "pure measure" of identifiability. With that caution in hand, we proceed to the positive contributions of these data. More stable sources of variation can be found when errors are studied vowel-by-vowel and when changes in discriminability are measured locally. Pooled error scores (as depicted in Figure 1) tend to obscure systematic differences among the vowels and, in this case, they mask some important similarities between the two tasks.

In Figure 2, percent misses on the monitoring task are presented for each vowel category. There was considerable variation among the nine vowels in their susceptibility to confusion and in their response to consonantal context. It is interesting to note that two of the short-long vowel pairs, the close vowels /I-i/ and /ʌ-u/, showed somewhat higher error rates in consonantal context than in isolation. In contrast, the open vowel pairs, /ɛ-æ/ and /ʌ-ɑ/, showed consistently lower error rates in consonantal context.

The differential effect of context on open and close vowels is presented more clearly in Figure 3. The bars on the left present pooled misses for the open-vowel pairs (/ɛ-æ, ʌ-ɑ/) and close-vowel pairs (/I-i, ʊ-u/) in the monitoring task. (Errors for /ɔ/ have been excluded from this analysis; their effect on the pattern in Figure 3 is, in any case, minor.) The bars are arrayed to allow a comparison of errors with and without the /pVp/ context. Elimination of the consonantal context produced a large increase in errors on the open vowels, but a small decrease in errors on the close vowels. Apparently, the labial-stop environment plays a much stronger role in the identification of open vowels than it does for close vowels. A similar kind of interaction can be observed on the right-hand side of Figure 3, where errors on the identification task have been pooled across open and close vowels. Elimination of the consonantal context produced a very large increase in errors on the open vowels, and a less marked increase for the close vowels.

The error data in Figure 3 do not give a direct measure of identifiability, unaffected by response biases. To obtain more direct measures, we used a log-linear model of the error data to compute discriminability scores for the most error-prone vowels in the monitoring task. The results verified what is visually salient on the left-hand side of Figure 3: the open vowel pairs, /ɛ-æ/ and /ʌ-ɑ/, both showed significantly greater discriminability in consonantal context. These changes in

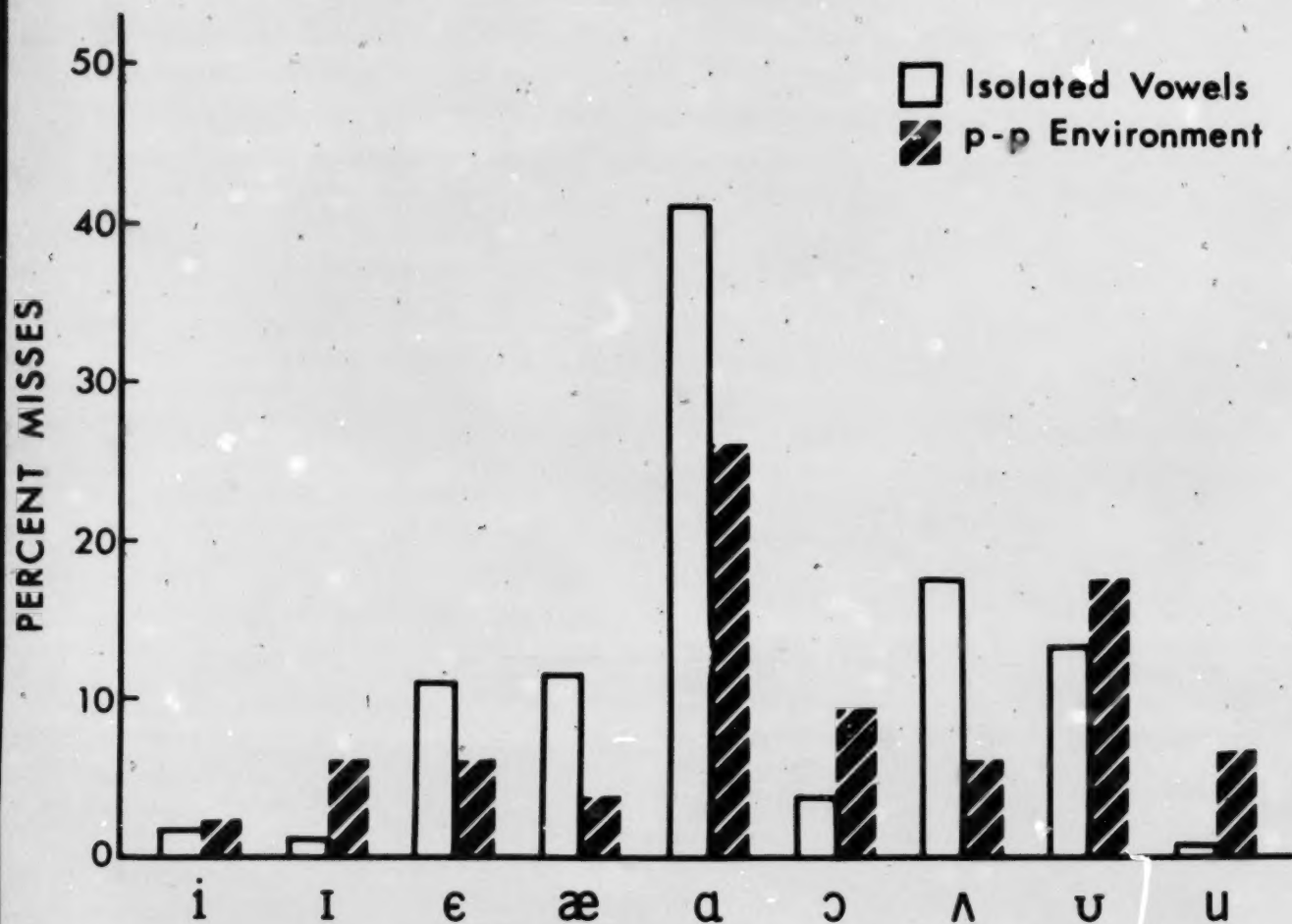


Figure 2. Mean percent misses on the nine vowels in the monitoring task. Data for isolated vowels and /pVp/ syllables are contrasted.

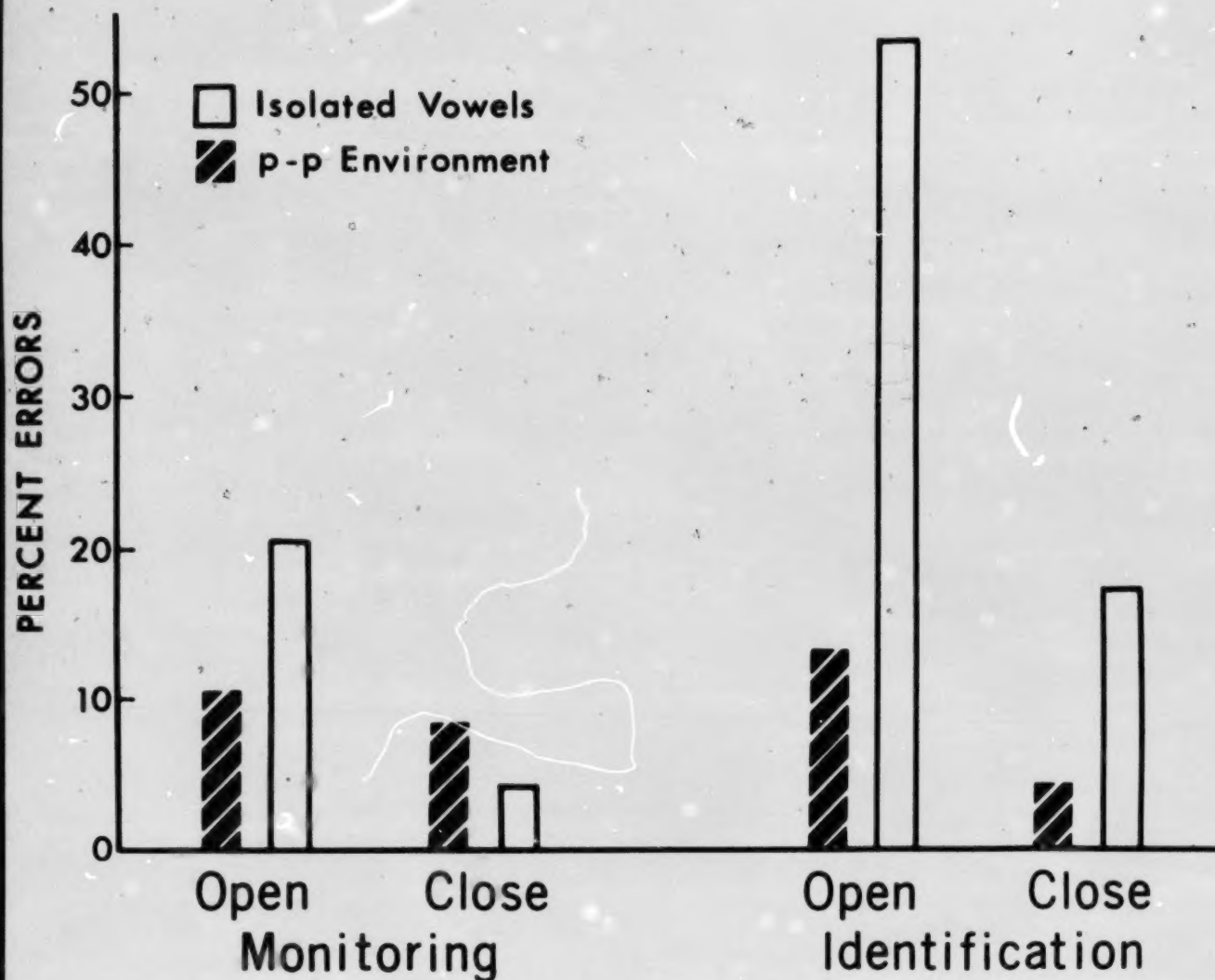


Figure 3. Mean percent errors in the monitoring and identification tasks, pooled over open (/ε, æ, ʌ, ɑ/) and close (/ɪ, ɪ, ʊ, u/) vowels.

discriminability were very similar to those observed in the identification task.

Thus, in spite of the manifest differences between the two tasks, a closer examination shows them to be in agreement on many points. Both sets of results demonstrate that a labial-stop environment affects the identification of specific subsets of vowels to differing degrees. This finding is only the starting point in our study of consonant-vowel contingencies; the acoustic and coarticulatory sources of these effects still need to be investigated in detail. We suspect that an adequate explanation for the effects will require analyzing acoustic syllable structure as a manifestation of (and specification of) coarticulatory interactions. For example, the contingency effects for labial stops could derive from a coarticulatory constraint that mainly affects the open vowels; specifically, the jaw movements for the open vowels are constrained because they must coarticulate with the jaw raising for the labial stop closure. As a result, it is possible that the timing relations holding over consonant and vowel are more tightly constrained for the open vowels than for the close vowels. This would presumably produce a sharper contrast in acoustic patterns and permit more accurate perceptual differentiation. The sensitivity of the short-long vowel pairs to consonantal context is one hint that timing relations may be an important contributor to the contingency effects.

One task for future research will be to measure the variability of timing relations in natural productions of CVC syllables and isolated vowels; the variability observed for specific consonant-vowel pairs may at least partially explain how the consonants influence perception of the vowels. This example serves to illustrate the direction of our ongoing research, which has as its overall objective to determine how specific coarticulatory contexts aid the perceptual definition of vowels in natural utterances.

REFERENCES

- Fowler, C. A., & Shankweiler, D. P. Identification of vowels in speech and nonspeech contexts. Journal of the Acoustical Society of America, 1978, 63, S4. (Abstract)
- Macchi, M. J. Are vowels in context really identified better than vowels in isolation? Journal of the Acoustical Society of America, 1979, 65, S6. (Abstract)
- Pisoni, D. B., Carrell, T. D., & Simnick, S. S. Does a listener need to recover the dynamic vocal tract gestures of a talker to recognize his vowels? Journal of the Acoustical Society of America, 1979, 65, S6. (Abstract)
- Shankweiler, D. P., Verbrugge, R. R., & Studdert-Kennedy, M. Insufficiency of the target for vowel perception. Haskins Laboratories Status Report on Speech Research, 1978, SR-55/56, 103-111.
- Strange, W., Edman, T. R., & Jenkins, J. J. Acoustic and phonological factors in vowel identification. Journal of Experimental Psychology: Human Perception and Performance, 1979, 5, 643-656.

- Strange, W., Jenkins, J. J., & Edman, T. R. Dynamic information specifies vowel identity. Journal of the Acoustical Society of America, 1978, 63, S5. (Abstract)
- Strange, W., Verbrugge, R. R., Shankweiler, D. P., & Edman, T. R. Consonant environment specifies vowel identity. Journal of the Acoustical Society of America, 1976, 60, 213-224.
- Tiffany, W. R. Nonrandom sources of variation in vowel quality. Journal of Speech and Hearing Research, 1959, 2, 305-317.
- Verbrugge, R. R., Strange, W., Shankweiler, D. P., & Edman, T. R. What information enables a listener to map a talker's vowel space? Journal of the Acoustical Society of America, 1976, 60, 198-212.

FOOTNOTE

¹By "error" we mean an occasion when the vowel identified by a listener differs from the vowel intended by the talker.

INFLUENCE OF PRECEDING FRICATIVE ON STOP CONSONANT PERCEPTION: SOME AMENDMENTS

Bruno H. Repp and Virginia A. Mann

Abstract. This paper presents two control experiments and one reanalysis of our earlier studies (Mann & Repp, 1979a) on the influence of [ʃ] vs. [s] on the perception of following [t] vs. [k]. While the basic finding of more [k] percepts following [s] was consistently replicated, the first control study failed to replicate the reduction in that effect when a syllable boundary intervened between fricative and stop (such an effect was reported in our earlier Experiment 1). The second control study showed that the relative amplitude of the fricative noise plays no important role in its effect on stop perception. However, this study and a reanalysis of our earlier Experiment 3 showed that--contrary to our earlier conclusion that the effect of the fricative on stop perception is primarily phonetic in nature--the spectral characteristics of the fricative noise do contribute significantly to the effect.

INTRODUCTION

In an earlier series of experiments (Mann & Repp, 1979a), we investigated the effect of a preceding fricative on the perceived place of stop consonant articulation. We preceded synthetic syllables from a [ta]-[ka] continuum with fricative noises appropriate to [ʃ] or [s] and showed that more velar stops are perceived in the context of [s]. Our earlier Experiment 1 demonstrated a decrease in the magnitude of this perceptual context effect with introduction of an additional vowel before the noise, which permitted a subjective syllable boundary after the fricative. Thus, instructions to place a syllable boundary between fricative and stop were confounded with the addition of a vowel preceding the fricative. Although it seemed unlikely that this initial vowel by itself should have been responsible for a reduction in the context effect, it was necessary to rule out that possibility by varying only instructions, holding the stimulus structure constant. This was the purpose of the following experiment.

EXPERIMENT 1

Method

Subjects. Ten subjects participated. They included eight paid volunteers, a research assistant, and the first author.

Acknowledgment. This research was supported by NICHD Grant HD01994 and BRS Grant RR05596 to the Haskins Laboratories, and by NICHD Postdoctoral Fellowship HD05677 to Virginia Mann.

[HASKINS LABORATORIES: Status Report on Speech Research SR-62 (1980)]

Stimuli. A new, improved [ta]-[ka] continuum was created on the OVE IIIc synthesizer at Haskins Laboratories. This continuum consisted of seven stimuli differentiated only by the F3 transition. The F3 onsets ranged from 2700 Hz to 2100 Hz in 100-Hz steps (plus or minus up to 10 Hz). F1 and F2 started at 285 and 1770 Hz, respectively. The steady-state frequencies of the three lowest formants were 771, 1233, and 2520 Hz. All formant transitions were stepwise-linear and 50 msec in duration. The specified amplitude rise-time was likewise 50 msec. Stimulus duration was 250 msec, and fundamental frequency fell linearly from 110 to 80 Hz. Informal listening indicated these stimuli sounded more natural and were easier to identify than the CV stimuli used in our earlier experiments.

The two synthetic fricative noises, appropriate for [ʃ] and [s], had the same spectral specifications as those used in our earlier experiments, but their duration was 200 msec, they had a triangular amplitude contour that rose during the first 150 msec and fell during the last 50 msec, and they were of approximately equal amplitude. Their amplitude was approximately 12 dB below the steady-state [a].

VFCV stimuli were assembled by following an initial, steady-state [i] vowel of 120 msec duration with either of the two fricative noises, a 75-msec gap, and one of the seven stimuli from the [ta]-[ka] continuum. Thus, there were 14 stimuli which were repeated ten times in random order, with interstimulus intervals of 3 sec, and 6 sec after each block of 14. The stimuli were digitized at 10 kHz before recording.

Procedure. Each subject listened twice to the same stimulus tape, once with "V-FCV instructions" and once with "VF-CV instructions". Under V-FCV instructions, subjects attempted to place a syllable boundary before the fricative and labeled the stops as "t" and "k". Under VF-CV instructions, they attempted to place the syllable boundary after the fricative and labeled the stops as "d" and "g". In each case, identification of the fricatives as "sh" or "s" was also required. The order of instructions was counterbalanced across subjects.

That the subjects were at least moderately successful in carrying out these instructions was indicated by some listeners' expressions of surprise when, on completion of the experiment, they learned that the same stimuli had been presented twice. Two subjects--the first author and his research assistant--were informed of this fact in advance; nonetheless, they found it possible to maintain a V-FCV or VF-CV "mode" of listening without difficulty and without involuntarily switching back and forth between the two modes.

Results and Discussion

The pooled results of the ten subjects are shown in Figure 1. The basic context effect--fewer alveolar stop responses following [s]--was replicated, $F(1,9) = 42.8$, $p < .001$. However, there was no indication of any effect of instructions, $F(1,9) = 0.2$. Even the two experienced listeners, who were convinced that they carried out the instructions successfully, both showed a slightly larger context effect in the VC-CV condition. Exclusion of trials on which the fricative was misidentified (3.0 percent) did not make any difference.

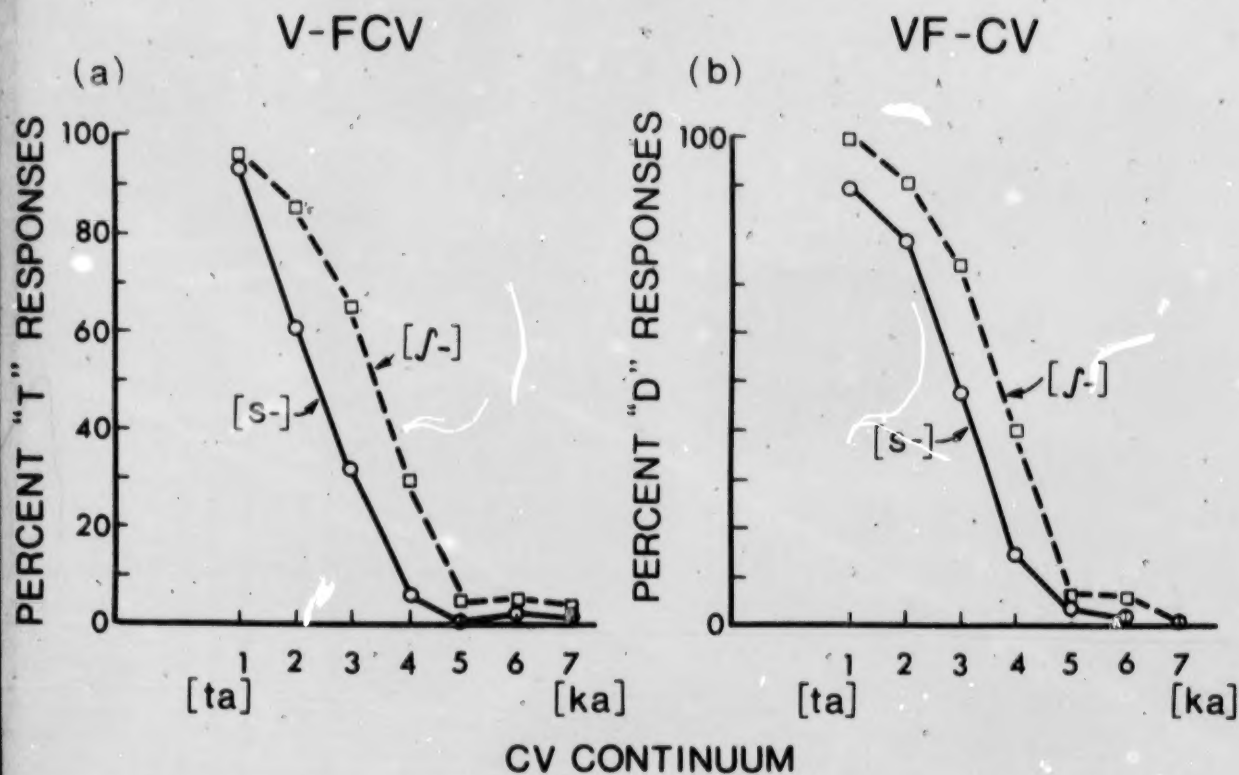


Figure 1. Effect of syllable boundary instructions (V-FCV vs. VF-CV) on the influence of fricatives on stop identification.

Thus, this well-controlled experiment failed to replicate the "syllable boundary effect" of our earlier Experiment 1. By default, that effect must be ascribed to the presence vs. absence of an initial vowel (VFCV vs. FCV). This interpretation is given some support by a comparison of the present results with those of Experiment 2, described below. In that experiment, the same subjects listened to stimulus sequences that included the present stimuli with the initial [i] portion removed. Indeed, the context effect was larger in FCV than in VFCV stimuli, although not significantly so. Thus, it seems that the context effect is sensitive to stimulus components preceding the fricative noise, but not to the cognitive variable of subjective syllable division.

EXPERIMENT 2

Experiment 2 had two purposes: (1) to investigate whether changes in fricative-noise spectrum within fricative categories would change the magnitude of the context effect, and (2) whether such within-category effects, if obtained, would diminish as the amplitude of the fricative noises is reduced. The presence of within-category effects would support the hypothesis that the context effect is due, in part, to auditory contrast and/or noise-offset cues to place of stop articulation (see Mann & Repp, 1979a), and sensitivity of these effects to noise amplitude would further point towards an auditory origin. Since fricative noise amplitudes had varied considerably among our earlier experiments, a controlled study of this factor was called for in any case.

Method

The subjects of Experiment 1 continued on to Experiment 2 in the same session. The stimuli were nearly the same, with three differences: (1) There were no initial vowels (i.e., the stimuli were of the FCV type); (2) instead of two fricative noises, there were four; (3) the noises existed in two versions, one "normal" (about 12 dB below the steady-state vocalic portion, as in Exp. 1), the other attenuated by 10 dB. The four noises were stimuli 1, 3, 7, and 9 from a nine-member [ʃ]-[s] continuum (see Mann & Repp, 1979b). Stimuli 1 and 3 sounded [ʃ]-like, stimuli 7 and 9 sounded [s]-like; thus, within-category effects could be assessed by comparing 1 with 3, and 7 with 9. Between-category effects, although confounded with a larger spectral difference, could be seen by comparing 1 and 3 with 7 and 9.

Two stimulus tapes were prepared, one with the normal noises and one with the attenuated noises. Each tape contained 10 random sequences of 28 stimuli, resulting from the combination of the four fricative noises with the seven CV stimuli, separated by 75 msec of silence. The order of the two tapes was counterbalanced across subjects.

Results

Figure 2 shows the stop identification results, with a summary of the data in the right-hand panel. As expected, there was a large effect of fricative category (fricative noises 1 and 3 vs. 7 and 9), $F(1,9) = 50.5$, $p < .001$. In addition, however, there was a significant within-category effect of fricative spectrum (fricative noises 1 and 7 vs. 3 and 9), $F(1,9) = 11.5$, $p <$

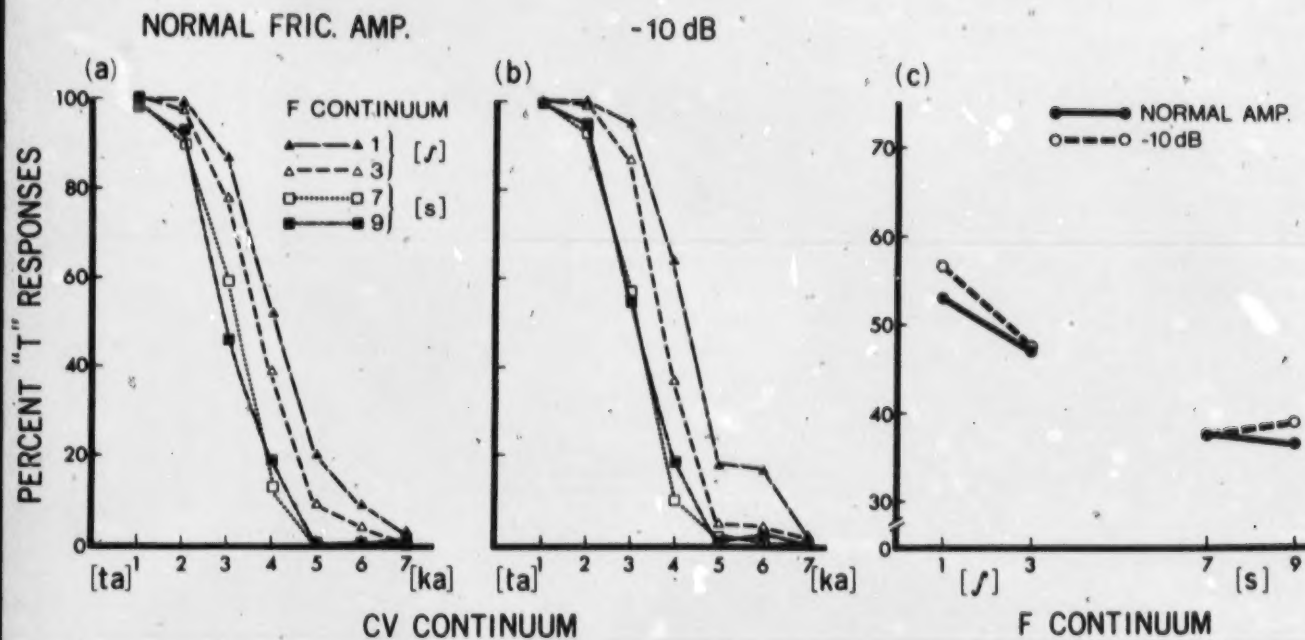


Figure 2. Effect of four different fricative noises on stop identification, for two different fricative noise amplitudes (panels a, b), summarized in panel c.

.01, but this effect was confined to the "sh" category (fricative noises 1 vs. 3), as indicated by a significant interaction, $F(1,9) = 10.0$, $p < .02$. Fricative noise amplitude had no effect at all; in fact, the effect of noise spectrum within the "sh" category was slightly larger for the attenuated noises.

The fricative noises were quite consistently identified as either "sh" (stimuli 1 and 3) or "s" (stimuli 7 and 9), regardless of amplitude, although "errors" were less frequent for the endpoint tokens. The error percentages for stimuli 1, 3, 7, 9 were 0, 4.1, 4.9, 1.5 at the normal amplitude and 1.9, 4.9, 2.4, 1.0 at the lower amplitude. Interestingly, [ʃ]-like stimuli were more often mislabeled as "s" in conjunction with a "k" response (2.1 percent) than in conjunction with a "t" response (0.6 percent), while the reverse was true for mislabelings of [s]-like stimuli as "sh" (0.8 vs. 1.5 percent). Exclusion of these errors from the data did not change the pattern of results.

Discussion

The results of this experiment suggest that the effect of the fricative on stop perception is, at least in part, a continuous function of fricative noise spectrum. Why effects of fricative spectrum were confined to the "sh" category is not clear at this point. Extrapolating from the trend within the "sh" category (Fig. 2c), it seems possible that the fricative context effect is entirely continuous and independent of perceived fricative category. However, the present experiment did not include any ambiguous fricative noises, which are needed to examine that issue. Experiment 3 in Mann and Repp (1979a) did precisely that and indicated that there is a large categorical component to the effect: "t" responses are less frequent when a given ambiguous noise is labeled "s" than when it is labeled "sh". This categorical effect persisted in our reanalysis of those data (see below).

Our failure to find any effects of noise amplitude in the present study is reassuring as far as amplitude variability in earlier experiments is concerned. However, it does not have any theoretical implications, as a positive finding might have had. The error pattern suggested that the stop may have a reciprocal effect on fricative identification. This possibility is further pursued in Experiment 3.

Before proceeding to describe that experiment, we should mention that we have some relevant data from another study that conflict with the present results in that they showed a within-category effect for "s" but not for "sh". In that study, (conducted for a different purpose and described in detail as Experiment 2 in Mann & Repp, 1979b), noises drawn from a nine-member [ʃ]-[s] continuum were followed by a variable silent interval and one of two fixed CV portions, called [ta] and [tu]. Since the formant transitions in the CV portions had not been chosen to be optimal for alveolar stops, and since the nine subjects were given the option to identify the stops as either "t" or "k", a large number of "k" responses was obtained, especially to stimuli containing [tu]. Not surprisingly, "k" responses were considerably more frequent following [s]-like noises than following [ʃ]-like noises. More detailed examination of the data showed that "k" responses remained at a constant level as long as the fricative was identified as "sh", but increased rapidly as noises identified as "s" became more and more [s]-like. In

conjunction with Experiment 2, these results show that within-category effects of fricative spectrum may differ quite radically from study to study—an impression that Experiment 3 confirmed.

EXPERIMENT 3

Experiment 3 examined the nature of within-category effects of fricative noise spectrum on stop identification and, in addition, tested whether the way a given fricative noise is labeled affects identification of the following stop. To that end, this study included several fricative noises ambiguous between [ʃ] and [s]. Thus it tested whether, in addition to the spectral effects demonstrated in Experiment 2, there is a categorical component to the context effect.

This experiment was previously reported by us as Experiment 3 in Mann and Repp (1979a). However, we subsequently discovered that our data analysis had used an incorrect procedure. Reanalysis produced a somewhat different pattern of results, so that our conclusions had to be modified. We also extracted some additional information not previously presented. Our present, complete account of this experiment supersedes that given in our earlier paper.

Method

Subjects. Ten adults served as subjects. They included eight paid volunteers and both authors.

Stimuli. The 25 FCV stimuli used in this experiment were formed by pairing each of 5 fricative noises with each of 5 CV stimuli, separated by a constant 75-msec period of silence. The stimulus components were similar to those employed in Experiment 2 of Mann & Repp (1979a). The fricative noises included the two unambiguous ones employed earlier and three noises ambiguous between [ʃ] and [s]. The ambiguity of these three additional noise stimuli was known from an earlier study of the [ʃ]-[s] distinction (Mann & Repp, 1979b); they were stimuli 4, 5, and 6 from a nine-member noise continuum. The CV stimuli were drawn from a [ta]-[ka] continuum (Mann & Repp, 1979a). They included the two endpoint stimuli and the three stimuli (4, 5, and 6) most ambiguous between [ta] and [ka].

Five randomized sequences of 55 stimuli were recorded directly from the synthesizer. Within each sequence stimuli that contained two unambiguous components were presented once, stimuli that contained one ambiguous component were presented twice, and those that contained two ambiguous components were presented three times.

Procedure. Each subject participated in a single 1-hour session, in which the test tape was presented twice. Thus, each subject gave a total of 30 responses to stimuli in which both components were (more or less) ambiguous. The task was to identify the fricative-stop cluster as "st", "sk", "sht", or "shk".

Results

The results for stop consonant identification are shown in Figure 3. The left panel displays standard labeling functions for the CV stimuli, separately for each preceding fricative noise. It is obvious that stop identification was as strongly affected by the preceding fricative as in the previous studies, $F(4,36) = 13.2$, $p < .0005$. The frequency of "t" responses declined steadily as the fricative noise became more [s]-like. To examine whether that decline can be accounted for entirely by the changes in noise spectrum, we turn to the right panel of Figure 3, which displays the overall percentages of "t" responses contingent on whether the fricative was identified as "sh" or "s". (Contingent percentages were calculated for each individual subject and stimulus combination before averaging; our failure to do this was responsible for the artifacts in our earlier analysis.) There we see that, although "t" responses declined within each fricative category, there was a sizeable break at noise stimulus 5, the one most consistently ambiguous between [ʃ] and [s]: When this particular noise was called "sh", "t" responses were more likely to follow than when it was called "s". This categorical effect was shown by seven of the ten listeners and, due to the very large effects shown by some subjects, was significant, $F(1,9) = 7.0$, $p < .05$.

In contrast to Experiment 2, the present data show declines in "t" responses within both fricative categories. In fact, the decline within the "s" category (not obtained in Exp. 2) was more consistent across subjects, $F(2,18) = 9.3$, $p < .005$; that within the "sh" category did not reach conventional levels of significance due to large individual differences, $F(2,18) = 3.2$, $p < .10$.

These results demonstrate that, for the majority of listeners, the fricative context effect has both a categorical and a continuous component. However, there are some listeners who seem to lack the categorical component, whereas others show an effect that is almost entirely categorical.

Examination of the fricative identification results revealed an effect of the CV portion, $F(4,36) = 6.7$, $p < .0005$. As Figure 4a shows, however, this effect was very small compared to that of the fricative on stop identification; it was primarily due to CV stimulus 1, which led to somewhat more "sh" responses than the other CV portions. This reciprocal effect exhibits the same contingencies as that of the fricative on stop identification: "sh" goes with "t", and "s" with "k". To see whether there is any categorical component to the reciprocal context effect, fricative identification was analyzed contingent on stop identification. The results are shown in Figure 4b. There were two consistently ambiguous CV stimuli (4 and 5), giving a good opportunity to observe categorical effects. However, the effects obtained were rather small (due in part to the small size of the total effect) and fell short of significance, $F(1,9) = 4.5$, $p < .10$. The decline in "sh" responses within the "t" category was significant, $F(2,18) = 7.9$, $p < .005$; there was no significant effect within the "k" category.

Discussion

Experiments 2 and 3 showed that the effect of the fricative on stop perception depends on perceived fricative category as well as on fricative

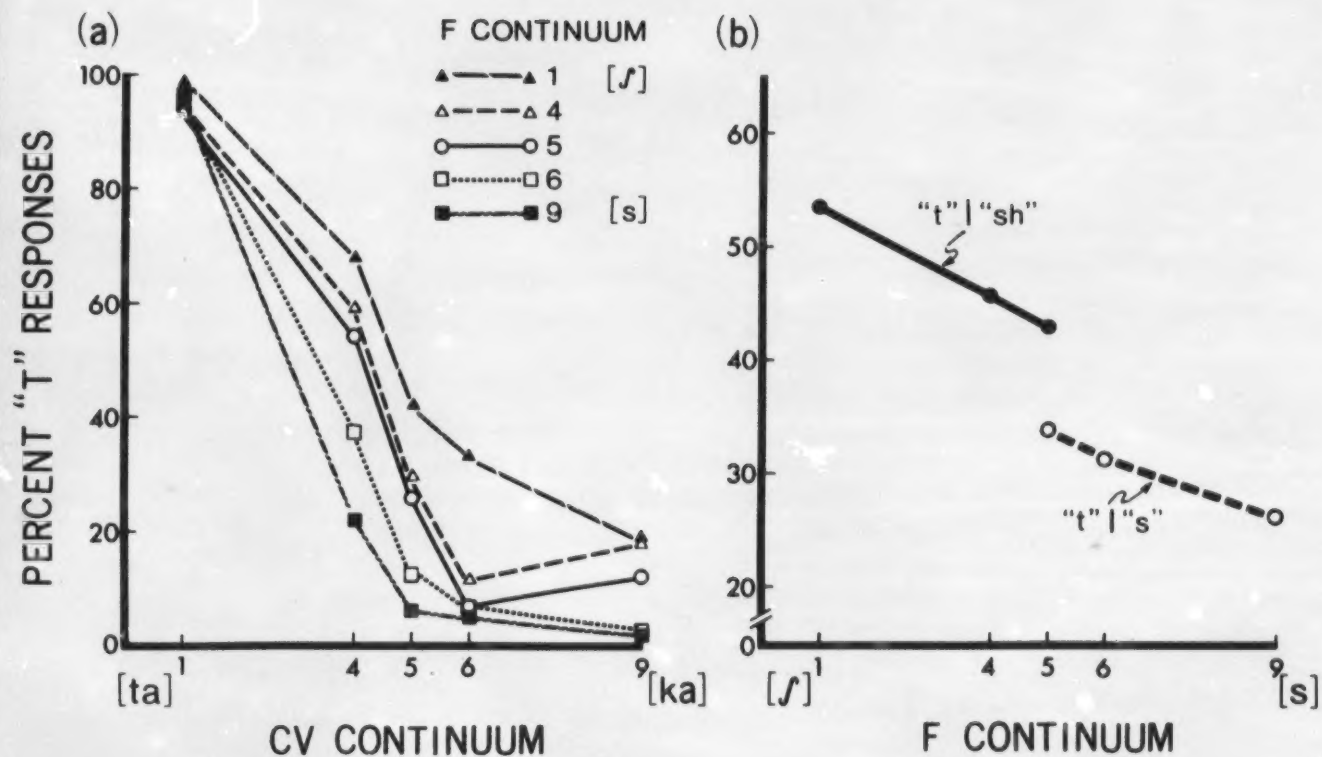


Figure 3. Effect of five different fricative noises on stop identification (panel a), and the continuous and categorical components of that effect (panel b).

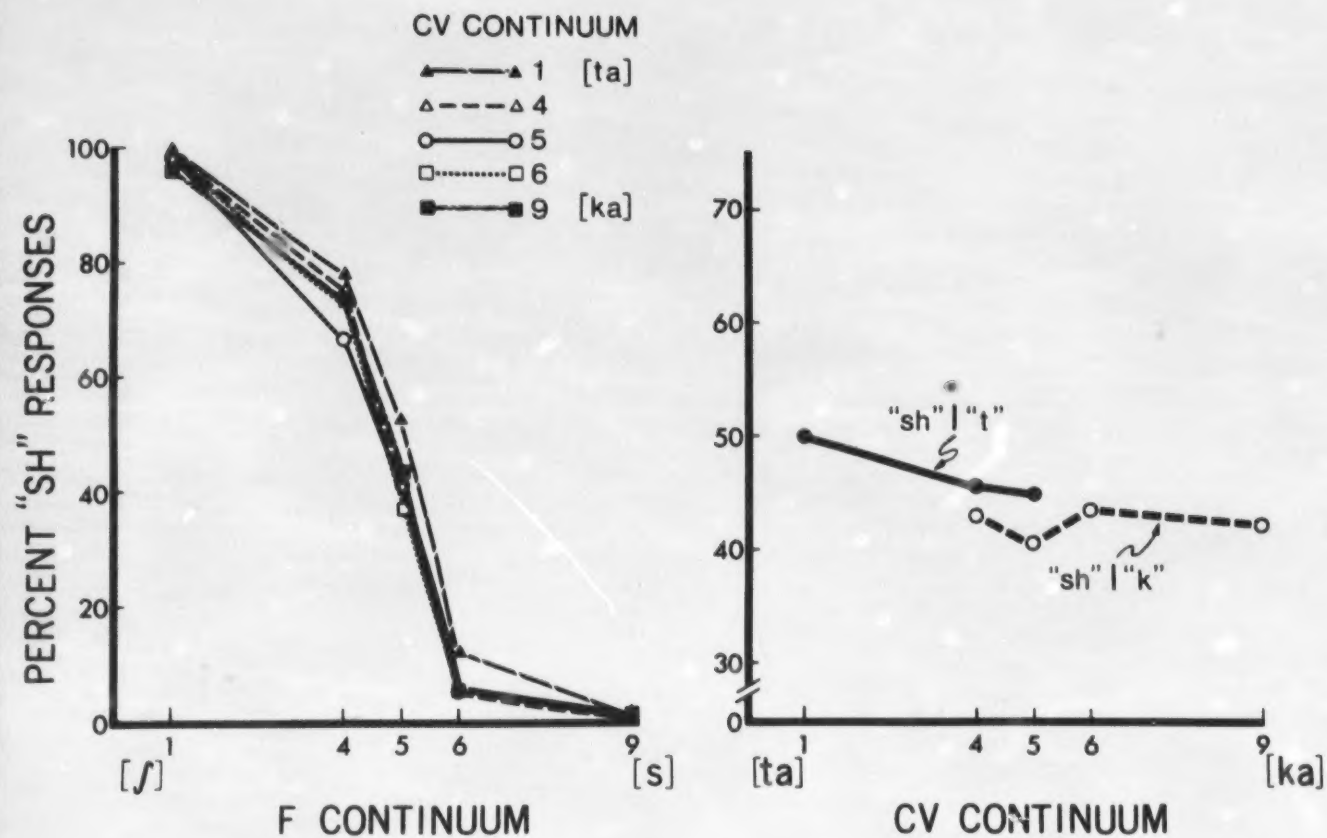


Figure 4. Effect of five different CV portions on fricative identification (panel a), and the continuous and categorical components of that effect (panel b).

noise spectrum. There is also a smaller, reciprocal effect of the stop on fricative perception, "s" responses being more frequent in conjunction with "k" responses, and--although the evidence is weak here--this effect may likewise have two components. The within-category effects of fricative noise spectrum on stop identification varied from study to study, even though highly similar stimuli were involved.

The finding that the context effect studied here is partially categorical and partially continuous makes it impossible to rule out any of the candidate theoretical explanations (discussed in detail by Mann & Repp, 1979a). Rather, we seem to need different explanations for the two components of the effect. Neither auditory contrast nor noise-offset cues to place of stop occlusion can explain the observed categorical effect; however, either of them may account for the somewhat elusive effects of noise spectrum within fricative categories. On the other hand, the categorical effect does fit a response bias hypothesis. Elsewhere, we have argued (Mann & Repp, 1979a) that the categorical effect does not derive from a simple response preference. We also have new evidence (Repp & Mann, Note 1) that subjects, when presented with isolated fricative noises and asked to guess the following stop, show no preference for "k" to go with "s" responses; this supports our argument.

It still seems plausible to us that the increased number of velar stop responses in the context of [s] may arise from listeners' implicit knowledge of certain variations (as yet not confirmed by acoustic measurements) in the transitions for velar and alveolar stops with the nature of the preceding fricative, reflecting a forward shift in the place of tongue-palate contact following [s]. Similarly, the effect of the stop on fricative identification may be due to the absence in our stimuli of appropriate transitions at the offset of the fricative noise. Since coarticulatory effects, such as the ones just referred to, are invariably assimilatory in character, perceptual compensation for them will favor contrasts; hence, listeners' preference for reporting combinations of fricatives and stops with opposite polarities on the place-of-articulation dimension. That the perceptual effect is partially categorical, partially continuous, may indicate that the compensatory mechanism operates both before and after phonetic category decisions. Thus, it seems that reference to articulation still provides the most parsimonious explanation of our results.

At present, we do not have any solid measurements demonstrating that there are, in fact, coarticulatory dependencies of stop consonants on preceding fricatives. However, we do have some perceptual evidence for the existence of such coarticulatory effects (Repp & Mann, Note 1). In the case of the reciprocal effects of stops on fricative identification, the existence of coarticulation in the form of spectral changes in the fricative noise is well-known, but we do not know whether these transitions affect the identification of the fricative itself. This possibility is not implausible, however, since the end of the noise seems to be more important for fricative identification than its beginning (Mann & Repp, 1979b).

It is possible that the perceptual effects investigated here reflect a very general principle of speech perception: a tendency to differentiate successive phonetic segments along the place-of-articulation dimension (cf. Repp, 1978). Such a tendency would compensate for the general assimilatory

nature of coarticulation, regardless of whether or not coarticulation occurs in a specific instance. We suspect, however, that the tie between speech perception and production is much more specific, with perceptual context effects occurring only when corresponding coarticulatory dependencies are in fact observed. Our present results may be an instance of such a relationship.

REFERENCE NOTE

1. Repp, B. H., & Mann, V. A. Perceptual assessment of fricative-stop coarticulation. Submitted for publication.

REFERENCES

- Mann, V. A., & Repp, B. H. Influence of preceding fricative on stop consonant perception. Haskins Laboratories Status Report on Speech Research, 1979, SR-59/60, 65-82. (a)
- Mann, V. A., & Repp, B. H. Effect of vocalic context on the [ʃ]-[s] distinction: I. Temporal effects. Haskins Laboratories Status Report on Speech Research, 1979, SR-59/60, 49-64. (b)
- Repp, B. H. Perceptual integration and differentiation of spectral cues for intervocalic stop consonants. Perception & Psychophysics, 1978, 24, 471-485.

SR-62 (1980)

II. PUBLICATIONS

III. APPENDIX

PUBLICATIONS

- Bell-Berti, F. Inter-articulator programming and clinical advice to "slow down." Journal of Childhood Communication Disorders, 1979, 3, 150-161.
- Bell-Berti, F. Book review of: "Time" in the production and the perception of speech, W. Barry & K. Kohler (Eds.). Journal of the Acoustical Society of America, 68, in press, 1980.
- Bell-Berti, F. A spatial-temporal model of velopharyngeal function. In Norman J. Lass (Ed.), Speech and language: Advances in basic research and practice, Vol. IV. New York: Academic Press, in press.
- Bell-Berti, F., Raphael, L. J., Pisoni, D. B., & Sawusch, J. R. Some relationships between speech production and perception. Phonetica, 1979, 36, 373-383.
- Drewnowski, A., & Healy, A. F. Missing -ing in reading: Letter detection errors or word endings. Journal of Verbal Learning and Verbal Behavior, 1980, 19, 247-262.
- Harris, K. S., & McGarr, N. S. Relationships between speech perception and speech production in normal hearing and hearing impaired subjects. In Proceedings of the Institute on Speech Assessment and Speech Improvement in the Hearing Impaired, National Technical Institute for the Deaf, Rochester, NY, June 22, 1979. Washington, D.C.: A.G. Bell Association, in press.
- Healy, A. F. Proofreading errors on the word the: New evidence on reading units. Journal of Experimental Psychology: Human Perception and Performance, 1980, 6, 46-57.
- Healy, A. F., & Levitt, A. G. Accessibility of the voicing distinction for learning phonological rules. Memory & Cognition, 1980, 8, 107-114.
- Katz, L., & Feldman, L. B. Linguistic coding: Comparison between a deep and a shallow orthography. In C. Perfetti & A. Lesgold (Eds.), Interactive processes in reading. Hillsdale, N.J.: Lawrence Erlbaum Assoc., in press.
- Löfqvist, A., & Yoshioka, H. Laryngeal activity in Swedish obstruent clusters. Journal of the Acoustical Society of America, in press.
- Mann, V. A., & Repp, B. H. Influence of vocalic context on perception of the [j]-[s] distinction. Perception & Psychophysics, in press.
- McGarr, N. S., & Harris, K. S. Articulatory control in a deaf speaker. In I. Hochberg, H. Levitt, & M. J. Osberger (Eds.), Speech of the hearing impaired: Research training and personnel preparation. Washington, D.C.: Alexander Graham Bell Association, in press.
- Niimi, S., Bell-Berti, F., & Harris, K. S. Dynamic aspects of velopharyngeal closure. Folia Phoniatrica, in press.

APPENDIX

DTIC (Defense Technical Information Center) and ERIC (Educational Resources Information Center) numbers:

<u>Status Report</u>		<u>DTIC</u>	<u>ERIC</u>
SR-21/22	January - June 1970	AD 719382	ED-044-679
SR-23	July - September 1970	AD 723586	ED-052-654
SR-24	October - December 1970	AD 727616	ED-052-653
SR-25/26	January - June 1971	AD 730013	ED-056-560
SR-27	July - September 1971	AD 749339	ED-071-533
SR-28	October - December 1971	AD 742140	ED-061-837
SR-29/30	January - June 1972	AD 750001	ED-071-484
SR-31/32	July - December 1972	AD 757954	ED-077-285
SR-33	January - March 1973	AD 762373	ED-081-263
SR-34	April - June 1973	AD 766178	ED-081-295
SR-35/36	July - December 1973	AD 774799	ED-094-444
SR-37/38	January - June 1974	AD 783548	ED-094-445
SR-39/40	July - December 1974	AD A007342	ED-102-633
SR-41	January - March 1975	AD A013325	ED-109-722
SR-42/43	April - September 1975	AD A018369	ED-117-770
SR-44	October - December 1975	AD A023059	ED-119-273
SR-45/46	January - June 1976	AD A026196	ED-123-678
SR-47	July - September 1976	AD A031789	ED-128-870
SR-48	October - December 1976	AD A036735	ED-135-028
SR-49	January - March 1977	AD A041460	ED-141-864
SR-50	April - June 1977	AD A044820	ED-144-138
SR-51/52	July - December 1977	AD A049215	ED-147-892
SR-53	January - March 1978	AD A055853	ED-155-760
SR-54	April - June 1978	AD A060448	ED-161-096
SR-55/56	July - December 1978	AD A065575	ED-166-757
SR-57	January - March 1979	AD A083179	ED-170-823
SR-58	April - June 1979	AD A077663	ED-178-967
SR-59/60	July - December 1979	AD A082034	ED-181-525
SR-61	January - March 1980	**	ED-185-636

AD numbers may be ordered from:

U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Road
Springfield, Virginia 22151

ED numbers may be ordered from:

ERIC Document Reproduction Service
Computer Microfilm International
Corp. (CMIC)
P.O. Box 190
Arlington, Virginia 22210

Haskins Laboratories Status Report on Speech Research is abstracted in Language and Behavior Abstracts, P.O. Box 22206, San Diego, California 92122.

**DDC and/or ERIC order numbers not yet assigned.